

MODELING DAYTIME AND NIGHTTIME POPULATION DISTRIBUTIONS IN  
PORTUGAL USING GEOGRAPHIC INFORMATION SYSTEMS

BY

C2007

Sérgio Manuel Carneiro Freire

Submitted to the graduate degree program in Geography and the  
Faculty of the Graduate School of the University of Kansas  
in partial fulfillment of the requirements for the degree of  
Master's of Arts

---

Dr. Johannes J. Feddema

---

Dr. Stephen L. Egbert

---

Dr. Jerome E. Dobson

Date Defended \_\_\_\_\_

The Thesis Committee for Sérgio Freire certifies  
That this is the approved Version of the following thesis:

MODELING DAYTIME AND NIGHTTIME POPULATION DISTRIBUTIONS IN  
PORTUGAL USING GEOGRAPHIC INFORMATION SYSTEMS

Committee:

---

Dr. Johannes J. Feddema

---

Dr. Stephen L. Egbert

---

Dr. Jerome E. Dobson

Date approved: \_\_\_\_\_

## ABSTRACT

Sérgio M. C. Freire

Department of Geography, December 2007

University of Kansas

Natural or man-made disasters (e.g., earthquakes, fires, toxic releases, terrorism, etc.) usually occur without warning and can affect large numbers of people. Census figures register where people reside and usually sleep, but when disaster strikes knowing where people are more likely to be at the time of the event can be invaluable information for adequate emergency response and evacuation planning. These data can also be useful for risk and consequence assessment or a variety of studies involving population, such as transportation planning, land planning, GeoMarketing, and health and environmental studies. Having this information in a GIS-usable raster format significantly increases its value and facilitates integration with other spatial datasets for analysis or modeling. The validity of the concept of ambient population for the desired purposes has been demonstrated by the recent development of global population distribution databases. However, the highest spatial resolution data available (30 arc seconds) for Portugal, though appropriate for use in major hazard events (e.g. volcanic eruptions, major earthquakes) which typically affect large areas, is too coarse for many practical uses in the country, especially at the local scale. Therefore, higher-resolution databases of daytime and nighttime population distributions are currently being developed by Los Alamos National Laboratory and Oak Ridge National Laboratory for the US, based on two different approaches.

This study concerns the development of fine-scale raster datasets of daytime and nighttime population distributions for two municipalities of Metropolitan Lisbon in Portugal, Cascais and Oeiras. Their combined population was 332,811 in 2001. The most recent census enumeration figures

and mobility statistics (2001) are combined with physiographic data, using areal interpolation in a dasymetric mapping approach. To model nighttime population, land use and land cover classes from different datasets are selected and combined with street centerlines to derive residential streets to which population is allocated in each census block group. The addresses of private businesses and public services (including health care facilities and schools) and respective workforce in each municipality are georeferenced to model the daytime worker population of 139,074, which is combined with a derived map of daytime residential population to estimate overall daytime population. Because reliable high-resolution results require accurate input data, a significant effort was devoted to verifying and improving input datasets, especially land use and land cover, street centerlines and business addresses.

Main results represent maximum daytime population and maximum nighttime residential population in 2001 for each 25-meter grid cell in the study area. Since the same spatial reference base was used to map population density, day and night distributions are directly comparable. Verification and validation procedures confirm that the approach suits the objectives. However, accuracy of results is mostly dependent on adequacy and quality of input data sets. Nevertheless, given the availability of input data sets for the whole country, it is possible to implement this methodology for other areas in Portugal.

**Keywords:** population density, ambient population, population modeling, emergency management, dasymetric mapping, areal interpolation, GIS, Cascais, Oeiras.

## TABLE OF CONTENTS

LIST OF FIGURES.....	vii
LIST OF TABLES.....	ix
LIST OF TABLES.....	ix
ACKNOWLEDGEMENTS.....	x
1. INTRODUCTION.....	1
1.1 Objectives .....	3
1.2 Significance of the Study .....	4
2. REVIEW OF THE LITERATURE .....	8
2.1 Approaches to Simulating Population Distributions.....	8
2.2 Nature of Population Data for Spatial Modeling .....	11
2.3 Population Modeling Methodologies .....	15
2.3.1 Areal Interpolation Methods .....	16
2.3.2 Areal Interpolation Without Ancillary Data .....	17
2.3.3. Areal Interpolation With Ancillary Data.....	19
2.3.4 Statistical Modeling Methods .....	26
3. METHODOLOGY .....	31
3.1 Study Area.....	31
3.2 Description of Software .....	34
3.3 Collection and Organization of Variables .....	36
3.3.1 Street Centerlines .....	37
3.3.2 Land Use and Land Cover .....	39
3.3.3 2001 Census Data .....	44
3.3.4 Workplaces and Employment .....	46

3.3.5 Commuting Statistics .....	46
3.4 Modeling the Population Distribution.....	47
3.4.1 Pre-processing of Data Sets .....	52
3.5 Nighttime Population Distribution.....	56
3.6. Daytime Population Distribution.....	62
3.7 Ambient Population Distribution.....	68
3.8 Verification and Validation .....	69
4. RESULTS AND DISCUSSION .....	71
4.1 Nighttime Population Distribution.....	71
4.2 Daytime Residential Population Distribution .....	74
4.3 Daytime Worker and Student Population Distribution .....	77
4.4 Daytime Population Distribution.....	80
4.5 Ambient Population Distribution.....	82
4.6 Verification and Validation .....	85
5. CONCLUSIONS .....	97
6. RECOMMENDATIONS FOR FURTHER STUDY .....	100
7. BIBLIOGRAPHY.....	102

## LIST OF FIGURES

Figure 1: Location of the study area in Portugal and in the Lisbon Metro Area. ....	33
Figure 2: Street centerlines obtained for the study area.....	38
Figure 3: The Land Cover Map of 1990 (COS'90) for the study area, converted to CLC nomenclature. See Table 5 for an explanation of the land cover classes.....	40
Figure 4: The CORINE Land Cover 2000 (CLC2000) map for the study area. See Table 5 for an explanation of the land cover classes .....	42
Figure 5: Density of resident population in the study area, by <i>secção</i> (block group).	45
Figure 6: Flowchart of main tasks involved in modeling the spatial and temporal distribution of population.....	51
Figure 7: Sample of original (A) and corrected (B) COS'90 LULC map.....	52
Figure 8: Flowchart of pre-processing of COS'90 LULC map. ....	53
Figure 9: Sample of original (A) and corrected (B) CLC2000 LULC map. ....	54
Figure 10: Street network in 2004 and in 2001. ....	55
Figure 11: Flowchart of GIS procedure for definition of residential land use polygons .....	59
Figure 12: Complete street network, residential polygons and residential streets in Oeiras. ....	60
Figure 13: Places of work and study georeferenced in Cascais and Oeiras.....	67
Figure 14: Grid of nighttime population distribution in Cascais.....	73
Figure 15: Grid of nighttime population distribution in Oeiras.....	74
Figure 16: Grid of daytime residential population distribution in Cascais. ....	76
Figure 17: Grid of daytime residential population distribution in Oeiras.....	77
Figure 18: Grid of daytime worker and student population distribution in Cascais....	79

Figure 19: Grid of daytime worker and student population distribution in Oeiras.....	80
Figure 20: Grid of daytime population distribution in Cascais. ....	81
Figure 21: Grid of daytime population distribution in Oeiras.....	82
Figure 22: Ambient population distribution in Cascais. ....	84
Figure 23: Ambient population distribution in Oeiras.....	85
Figure 24: Comparison of census vs. modeled population by block in Cascais. ....	89
Figure 25: Comparison of census vs. modeled population by block in Oeiras.....	89
Figure 26: Map of percentage error by census block in Cascais and Oeiras.....	91
Figure 27: Map of mean absolute percentage error (MAPE) by census block group in Cascais and Oeiras. ....	93
Figure 28: Map of count error by model cell for nighttime distribution in Cascais.....	94
Figure 29: Map of count error by model cell for nighttime distribution in Oeiras. ....	95



## LIST OF TABLES

Table 1. Typology of methods for modeling population distribution.....	15
Table 2. Characteristics of global gridded databases of population density .....	29
Table 3. Population and growth rates in census years, 1981-2001.....	34
Table 4. Main input datasets used for modeling nighttime and daytime population..	37
Table 5. Standard CORINE Land Cover nomenclature.....	42
Table 6. Results of correlation analysis of cell size in Oeiras.....	50
Table 7. Summary characteristics of population modeling approach .....	57
Table 8. Resident population and figures calculated from O/D matrix in the study area, 2001.....	63
Table 9. Characteristics of census blocks and block groups in Cascais and Oeiras.	86
Table 10. Population statistics by block for Cascais and Oeiras. ....	87
Table 11. Overall accuracy measures for Cascais and Oeiras.....	87
Table 12. Statistical characterization of percent error by census block in the study area.....	92

## ACKNOWLEDGEMENTS

Fortunately no research work is entirely conducted in isolation, even when the journey is long and often solitary. I would like to thank Johan Feddema for his guidance, for offering me a research assistantship, and for boldly believing this could be done. Special thanks go to my thesis committee members, Steve Egbert and Jerry Dobson, whose suggestions greatly improved this work - Steve, your combination of teaching excellence and personal qualities is truly unique. I should also thank Johnathan Rush and Mike Brown of Los Alamos National Laboratory, on whose work this project was based, for their input in clarifying finer details. Being a student at the University of Kansas was a very rewarding experience, and a sincere word of appreciation goes to everyone I was fortunate to meet at the KU Geography Department. I am deeply grateful to George McCleary, Bev Koerner, and especially to former chairman Bob McColl who so generously welcomed me to Lindley Hall. Thanks are also due to Steve Driever of UMKC for his support and for offering me a chance to practice and develop my research skills.

A special word of gratitude goes to my mother and sister for first suggesting that I join the KU Graduate Program and supporting me in that decision. I also thank my friends and former colleagues Hugo Carrão and António Nunes for their wise opinions, aid with finer points and often-needed words of encouragement. I acknowledge the generosity of GeoPoint Lda., which made many things easier. I thank my great friends Ivan and Andrea Damjanov and everyone who directly or indirectly contributed to me graduating today.

With this work I tried to contribute something useful and that could make a difference for the better, and in that sense the present manuscript or degree cannot be an end in itself. While deficiencies and shortcomings in this work can only be attributed to me, whatever good comes out of it is certainly

dedicated to my dear wife Ana for her ever-present encouragement, patience, and love.

## 1. INTRODUCTION

Natural or man-made disasters, such as natural disasters, technological accidents, and terrorism usually occur without warning and can potentially affect people from a local to continental scale. Accurately estimating population exposure is recognized as a key component of catastrophe loss modeling, one element of effective risk management (FEMA, 2004; Chen et al., 2005; NRC, 2007). Despite recent efforts by Dobson (2002; 2003; 2007) at devising a technique that could be employed in real-time once a disaster occurs, for planning and simulation purposes and to ensure adequate timely response, population distribution information should be produced and made available beforehand whenever possible. Also, such data sets can be useful for virtually any application involving the spatial distribution of people if they are produced at appropriate, application-specific, spatial and temporal scales (Sutton et al., 2003).

Having this information in a Geographic information System (GIS)-usable raster format significantly increases its value by facilitating integration with other spatial datasets for analysis or modeling. Although efforts to rasterize population distributions predate the development of most current commercial GIS (Balk et al., 2006), increased availability of digital spatial data combined with the improved analysis capabilities in GIS have allowed for the development of several global population distribution databases (Tobler et al.,

1995; Goldewijk and Battjes, 1997; Dobson et al., 2000). However, their spatial detail is still insufficient to adequately support analysis at the local level and to distinguish between daytime and nighttime population distributions.

Population distributions are not static in time, varying over daily, seasonal and long term time scales (Sutton et al., 2003). For disaster planning in urban areas, it is the daily population variation that is particularly important to be able to estimate the number and socioeconomic classes affected by an impact. For a given area, daytime population distribution is likely to differ from nighttime due to a number of human activities, such as work and leisure. In Portugal, existing population distribution maps and exposure studies are based on census data (e.g., Oliveira et al., 2005). Census figures register where people reside and usually sleep, but, when disaster strikes, knowing where people work or shop can be invaluable information for adequate emergency response and evacuation planning.

Population density seems like a simple measure, but the nature of human population counts as a spatial variable poses several challenges with respect to its accurate representation at a given spatio-temporal scale. While census data can be taken to reasonably approximate nighttime population counts, to estimate daytime population distribution usually other types of information from other sources are needed. For the purpose of data comparability between daytime and nighttime population distributions, it is

desirable that these datasets be available or converted to the same spatial reference base.

Modeling efforts by various authors have employed differing approaches and methodologies to address these challenges and represent the distribution of population in diverse regions and at varying scales.

### 1.1 Objectives

The main objective of this study is to develop and implement a data-driven model to map current daytime and nighttime distributions of population in Portugal at high spatial resolution, using readily available data sets and statistics. The model is tested in two municipalities of the Lisbon Metropolitan Area, Cascais and Oeiras, but since it relies on mostly publicly-available geographic data sets and statistics, this approach can be used to map population distribution in other areas of Portugal. The model also aims to approximate a representation of ambient population through the combination of daytime and nighttime distributions in a single measure. Dobson et al. (2000), in the context of the population dataset called LandScan, coined the concept of “ambient population” as a temporally averaged measure of population density that accounts for the *loci* of human activities (such as sleep, work, study, transportation, etc.) that may be more adequate for certain

applications (such as emergency response) than residence-based population density.

The secondary objectives of this project are to examine the adequacy of using residential streets information as a proxy for disaggregating census data at fine spatial scales, and to analyze the effect of grid cell size on the accuracy of this type of dasymetric map, following Eicher and Brewer's (2001) suggestion.

## 1.2 Significance of the Study

Despite the continuing interest in measuring and describing the spatial characteristics of human population and socioeconomic variables (e.g. by census bureaus, etc.), in terms of accuracy and spatial coverage these measures still lag behind the measurement and representation of physical variables describing the Earth's surface (Deichmann, 1996). Examples of such physical factors include accurate high resolution elevation datasets, and remote-sensing derived land use and land cover maps. This also holds true in Portugal where a single national population distribution database is non-existent among the many publicly-available digital spatial datasets, despite the country's pioneering role in developing a Spatial Data Infrastructure and repository (Julião, 2003).

This study originated from an interest in the development and applications of the LandScan Global Population Project (Dobson et al., 2000), and the realization that, despite its originality and intrinsic value, these data were not suitable for certain applications, especially those requiring higher temporal and spatial detail. Dobson (2002) acknowledges that “even finer resolutions are needed for many types of disasters”, namely those that can “impact areas as small as a neighborhood, city block, or single building”, and therefore the Oak Ridge National Laboratory (ORNL), based on the 2000 census, is developing LandScan USA as daytime and nighttime population surfaces at the higher resolution of 3 arc seconds (Bhaduri et al., 2002).

This project follows a model successfully developed and tested by Los Alamos National Laboratory (McPherson and Brown, 2003; McPherson et al., 2004; McPherson et al., 2006) to estimate daytime and nighttime population distributions in U.S. cities for emergency response activities. The Los Alamos approach is adapted and tested in Portugal using recently available spatial data sets and statistics to map diurnal and nocturnal distributions of population at high spatial resolution. It is expected that these maps will suit a variety of purposes requiring population information. However, the main intent of the project is to develop a population dataset to aid emergency management activities by improving daytime exposure estimates of populations potentially affected by natural or human-induced hazards. Other



possible uses for the data include spatial modeling, health and environmental studies (epidemiology, pollution, quality of life, etc.), GeoMarketing, transportation and urban planning, and education.

The significance of this study is based on the following points:

1. Results meet the need for more and better geographic socioeconomic databases by providing insight into the spatial and temporal distribution of population in the study area at resolutions previously unavailable; the high spatial resolution permits analysis at the local scale for which existing population databases are inadequate, while allowing for easy spatial aggregation to coarser cell sizes when appropriate.
2. Development of a methodology or model that is data-driven and mostly dependent on official census and statistics of population counts and distribution, as opposed to relying on empirical or heuristic weights which usually make model calibration difficult and uncertain; also, since input data exist for other municipalities in Portugal, population can be modeled for those areas as well.
3. The approach models both worker/student and residential components of daytime population, and yields potentially

valuable and previously unavailable intermediate data products such as residential streets and georeferenced workplaces.

4. Results can be used in the future to derive coefficients for dasymetric interpolation of nighttime and daytime population distributions for larger areas (e.g., whole metropolitan areas), especially in the daytime period.

## 2. REVIEW OF THE LITERATURE

### 2.1 Approaches to Simulating Population Distributions

Numerous efforts have attempted to portray population distribution on a regular raster grid. Deichmann (1996) refers to early examples such as Adams' (1968; see Deichmann, 1996) population density map for West Africa (which served mostly cartographic purposes), national population grids that have been produced for decades by the census offices of Japan and Sweden, and digital maps for individual countries produced by the US Census Bureau (Leddy, 1994). By the 1970's, the Oak Ridge National Laboratory was using computers to map the US population density for small areas (Haaland and Heath, 1974). Although attempts at creating gridded population surfaces predate the advent of the computer (Balk et al., 2006; DeMers, 1997), the last two decades have seen numerous studies and initiatives to further this methodology. Chief among the many factors that contribute to a renewed interest in population distribution methods and applications, are: a) the advancements of Geographic Information Systems (GIS) and GIS-based analysis and mapping, b) greater availability and integration of digital spatial data (digital maps, satellite imagery, etc), c) an increased awareness and interest in the socioeconomic dimension in environmental change studies,

and d) a need to improve emergency management by better estimating population exposure to risk.

This context has fostered the research and production of numerous population distribution databases at different spatial resolutions and having local to global coverage. The approaches behind these datasets range from heuristic (i.e., without modeling) to experimentation with several recent “intelligent” areal interpolation methods. Notably, the combination of areal interpolation research and dasymetric mapping applications resulted in added complexity (Fisher and Langford, 1996; Eicher and Brewer, 2001). Each effort usually represents a unique combination of: a specific purpose or goal, variables used, adopted methodology, scale of study, and envisioned application(s). Although the adoption of one or another of these inter-related criteria complicates the task of trying to systematize a fast-evolving field (as illustrated by the different designations used by different authors for the same techniques or its variants), most efforts reported in the GIS and remote sensing literature seem to fit the comprehensive typology suggested by Wu et al. (2005).

Regarding the purpose for modeling, there are two main reasons to estimate population distribution surfaces. The purpose to a great extent determines the adopted interpolation methodology. From this perspective the processes of modeling population distribution can be grouped under the following general approaches (Wu et al., 2005):

1. An approach mostly concerned with refining the population distribution for a given area by realistically redistributing existing population counts, from census or other sources. To this effect, interpolation or disaggregation techniques are usually applied in what may be considered a “top-down” approach to modeling. The present study fits into this category.
2. A statistical modeling approach is usually employed when the main purpose is to estimate the total population for an area at a specific time, normally when an official or reliable count is unavailable. The goal is accomplished by inferring or applying a statistical relationship between population and other related variables, such as city size or land use.

Deichmann (1996) also divides approaches to this problem into two categories, areal interpolation and surface modeling. However, that author argues that surface modeling can also be seen as a special case of spatial interpolation when it is used as an intermediate step in addressing the problem of incompatible zonal systems.

More recently a new set of approaches is emerging that is aimed at directly estimating the population distribution from which total counts can subsequently be derived for any zoning. Such a “bottom-up” approach was experimented and compared by Rabbani (2007), who estimated ambient

population for a municipality in Brazil by deriving building occupancy coefficients, and combined this with a buildings database.

## 2.2 Nature of Population Data for Spatial Modeling

Population figures are among the most basic socioeconomic indicators. However, as with any geographic variable, the way demographic data are collected and made available have implications for its manipulation and representation. The complex nature of population data poses several challenges for adequately modeling and analyzing its distribution in space and time. Goodchild et al. (1993) and Deichmann (1996) mention a number of characteristics and difficulties that make representation of these data problematic, including:

1. That, despite being collected for individual households, owing to confidentiality and data volume requirements, census counts are reported as aggregate figures for a set of larger collection units. Therefore geographic population databases have originally two basic components: a spatial reference system of contiguous polygons and respective attribute data. When the reporting units differ from the ones used for analysis, a method needs to be employed to transform census data between these

incompatible spatial units, potentially creating interpolation/extrapolation errors.

2. When the nature of reporting units is arbitrary, results of a spatial analysis become dependent on that configuration, an effect known as the modifiable areal unit problem (Openshaw, 1983). Data characteristics may also change with aggregation level – i.e., the concept of ecological fallacy, which suggests general population characteristics apply to individuals (Openshaw, 1984).
3. Census data are basically residence-based. However, population is not static, and its variation in space and time is scale-dependent (Sutton, 2003). Due to human activities and mobility, census data represent a specific situation that in reality may never occur. The distinction made in some national censuses between de jure (i.e., usually resident) and de facto (present) population is mostly ineffective at adequately addressing this issue. In addition to these issues, some applications require knowledge of non-residential population distributions. In particular, emergency management applications have suggested the need for temporal disaggregation of population distributions. This has been achieved mainly through the simulation of ambient population and segmentation of

population information into daytime and nighttime distributions. The concept of ambient population was first proposed by Dobson et al. (2000), and was heuristically implemented as a temporally averaged measure of population density that accounts for human activities. In this sense, ambient population corresponds to a temporal aggregation of population distribution, a compromise between daytime and nighttime distributions that strictly represents neither period. Sutton et al. (2003) suggest that smoothing of residential-based population grids by using a mean spatial filter can also approximate ambient population distribution. However, this requires knowledge of specific mobility patterns, assumes these are stable, and even then can only simulate displacement in the vicinity of residence.

Recently there has been more research on the temporal dimension of population distribution (e.g., Zandvliet and Dijst, 2004), and daytime and nighttime population datasets have been created for various areas at several resolutions: McPherson and Brown (2003) estimated daytime and nighttime population distributions in U.S. cities for emergency response activities at 250 meter spatial resolution, while the Oak Ridge National Laboratory is creating LandScan USA as daytime and



nighttime population surfaces for the year 2000 at the higher resolution of 3 arc seconds (Bhaduri et al., 2002); at the local level, Sleeter (2004) redistributed census population to a 30-meter grid in the San Francisco Bay region, USA, and Sleeter and Wood (2006) estimated day and night population densities at 10 meters for a coastal county in Oregon.

4. Population counts are discrete, ratio-level numeric data. Thus population density is inherently a discrete statistical surface, “one that occurs only as individuals with some difference in numbers per unit area” (DeMers, 1997, p. 258). Although its distribution may be approximated by a continuous surface (e.g., isopleth maps), this becomes impossible below a certain level of spatial resolution, which itself depends on density (Goodchild et al., 1993). Also, since population refers to specific positive quantities, methods used for accurate representation of population surfaces should satisfy the necessary (but not sufficient) volume preserving and nonnegativity requirements (Tobler, 1979). Tobler’s so-called “pynophylactic condition” implies that within one enumeration zone, the sum of all distributed individuals should match input totals for that zone.

### 2.3 Population Modeling Methodologies

The variety of modeling methodologies present in the literature also reflect the complex nature of population data as evidenced by the assumptions and operations made in order to modify the data for specific applications. Based on the many methodologies it becomes clear that there is no single best method for interpolating or rescaling the data, instead each method has its strengths and limitations for a desired purpose.

The typology of methods and corresponding techniques suggested by Wu et al. (2005) is summarized in Table 1.

Table 1. Typology of methods for modeling population distribution

Methods				Techniques
Areal interpolation	Without ancillary data	Point-based	Exact	distance-weighting, kriging, spline, finite difference
			Approximate	least squares, least squares fitting with splines, Fourier series models, power-series trend models
		Area-based		Areal weighting, Pycnophylactic interpolation
	With ancillary data	“Intelligent” interpolation		Control zones
		Dasymetric mapping		Limiting variables, Related variables
Statistical modeling				Correlation with urban extent, land use, dwellings, image pixel, and several physical and socio-economic variables

*Source: Modified from Wu et al. (2005)*

### 2.3.1 Areal Interpolation Methods

Cross-area estimation or areal interpolation (Goodchild and Lam, 1980) is primarily designed for transferring data between two sets of non-nesting spatial units, a process also designated by Goodchild et al. (1993) as spatial basis change. The two spatially incompatible data location arrays are usually termed source zone and target zone, corresponding to census enumeration units and grid cells, respectively, in the context of the current population interpolation project.

According to Wu et al. (2005), the quality of the resulting estimates seems to depend largely on a) the definition of source and target zones, b) the degree of generalization in the interpolation process and c) the characteristics of the partitioned surface, or the assumptions made regarding the homogeneity of population distribution in either source or target zones (Goodchild et al., 1993; Deichmann, 1996).

This approach can be divided into two categories, those that incorporate ancillary information as surrogate variables to aid the interpolation process, and those that do not use such ancillary information. However, most methods in the latter group are also capable of integrating surrogate variables.

### 2.3.2 Areal Interpolation Without Ancillary Data

#### 2.3.2.1 Point-Based Methods

Either point-based methods or area-based methods can be used when population data are the only input for interpolation. In point-based interpolation, local population counts are assigned to point locations whose distribution is assumed to be a summary of the distribution of the variable to be modeled (Martin and Bracken, 1991). With census data, point locations that represent each source zone value can be used to generate a population grid. Exact methods preserve the original point values and include most distance-weighting methods, kriging, spline functions, and finite difference methods. Inexact or approximate methods are concerned with determining an overall surface function at the expense of maintaining point values and include distance-weighted least squares, least squares fitting with splines, Fourier series models, and power-series trend models.

Other than their complexity, point-based methods entail significant problems, such as: a) source zone centroids, usually taken as point values for interpolation, are not true “sample locations” and may even fall outside of the representative source zone, which is a relevant concern since the results are particularly affected by the density and spatial arrangement of data points; b) the a priori assumptions made by interpolation methods concerning the

surface involved - these seldom fit complex geographical phenomena; and c) methods that do not meet the volume-preserving requirement, thus being fundamentally inadequate for modeling population density, especially high resolution problems where a smooth continuous surface does not match the discrete nature of the variable. Notable exceptions are the distance-decay function by Martin and Bracken (1991), the modified kernel-based function by Martin (1996), and the quadratic kernel function described in Silverman (1986) which is the base for the Kernel Density tool available in the ArcGIS 9.1 software (ESRI, ArcGIS Help).

#### 2.3.2.2 Area-Based Methods

An immediate advantage of area-based methods is their volume-preserving ability. In this category, the simplest implementation is the method known as areal weighting, whereby target zones are overlaid on source zones and the resulting proportions serve as a weight to linearly apportion population. However, this procedure assumes population density to be constant within each source zone, a condition usually violated in real census areas.

Proposing an alternative, Tobler's (1979) raster-based pycnophylactic (i.e., mass-preserving) interpolation tends to the other extreme by artificially imposing a smooth transition between adjacent zones to minimize the

curvature of the estimated surface. However, this can also lead to artificial errors because abrupt discontinuities do exist in the spatial distribution of population density at local scales, which make this approach and its variants less realistic and effective for such applications (Flowerdew and Green, 1992). In general, overlay methods produce better results for discontinuous surfaces while Tobler's method is preferable when smoothness is an actual surface feature.

### 2.3.3. Areal Interpolation With Ancillary Data

Since the distribution of population is related to other spatial variables, namely land use, terrain slope, and transportation networks, such ancillary information can be used to aid in the interpolation process.

Some area-based approaches aimed at overcoming the problem caused by the assumption of homogenous source zones have been labeled as "intelligent" interpolation methods. Among these are studies by Flowerdew and Green (1992) and Goodchild et al. (1993) where so-called "control zones", believed to have homogenous densities, are used as a third set of areal weights to estimate population values. These authors demonstrated that the use of even subjective or imprecise geographic information can significantly improve the accuracy of results.

In Portugal, Néry et al. (2007) built on the efforts of Vidal et al. (2001) and Goodchild et al. (1993) to experiment with zonal interpolation methods for socio-economic statistics. They demonstrated their usefulness for spatial disaggregation of census data, and show that improvements are obtained by including land cover information as ancillary data.

#### 2.3.3.1 Dasymetric Interpolation

The fact that socio-economic data are usually reported as aggregate figures for a set of collection units explains why choropleth mapping is the most popular and simple technique for portraying area-based data such as population counts. Choropleth maps depict a statistical surface using area symbols that coincide with the data collection regions (Robinson et al., 1995). The problem with this type of mapping is that it conveys an idea of uniform distribution of the variable within each reporting unit. In the context of population census mapping, because mapping zones are enumeration zones, there is an exhaustive coverage of space even in areas of the map where population is not present.

Aiming to overcome this limitation of choropleth mapping, dasymetric mapping can be defined as a cartographic technique that allows the depiction of quantitative areal data using boundaries from additional ancillary information (objective or subjective) that divide the mapped area into zones of

relative homogeneity – dasymetric zones. Dating from the nineteenth century, this technique, popularized in the U.S. by Wright (1936), was originally used for population mapping (Eicher and Brewer, 2001, citing McCleary, 1969; 1984). According to Maantay et al. (2007), the earliest known example of its application is Scrope's 1833 classed population density map of the world, but the Russian geographer Semenov Tyan-Shansky (1827-1914) is usually credited for inventing the dasymetric map, although this issue is still a matter of academic debate. The technique has recently experienced a revival due to advances offered by GIS (DeMers, 1997). Other practical problems where dasymetric mapping has been successfully applied include crime mapping and catastrophe loss estimation. For example, Poulsen and Kennedy (2004) used dasymetric mapping to disaggregate the spatial distribution of residential burglary in Massachusetts, USA, based on land use and housing data. Chen et al. (2004) employ the technique to refine housing distributions to map exposure estimates for catastrophe loss estimation in Sydney, Australia. The authors recognize the potential benefit to risk assessment by employing exposure data to finer areal units, as opposed to the previous method that omitted spatial disparity in loss estimation models.

Dasymetric mapping is frequently applied without being explicitly acknowledged in the GIS and remote sensing literature (Robinson et al., 1995; DeMers, 1997), falling generally into the generic modeling category. Although dasymetric mapping largely predates studies in areal interpolation



methods, the overlap of dasymetric mapping applications and areal interpolation research has resulted in added complexity and ambiguity for the field (Eicher and Brewer, 2001). Chrisman (2002) argues that dasymetric mapping is in fact a form of areal interpolation and therefore suggests the use of the term *dasymetric interpolation* instead of the more commonly used designation.

Recent applications of dasymetric mapping or interpolation using GIS, both as raster and vector-based methods, include Eicher and Brewer (2001), Harvey (2002a), Mennis (2003), Sleeter (2004), Mennis and Hultgren (2006a, 2006b), and Langford (2006; 2007). Raster methods have not been popular for dasymetric mapping, but their effectiveness combined with ease of implementation and use will probably lead to their more frequent adoption (Eicher and Brewer, 2001). Raster based dasymetric mapping with adequate resolution can be effective at bridging the gap between visualization-oriented choropleth maps and analysis-oriented areal interpolation.

Dasymetric mapping allows ancillary data related to the mapped variable to be explored in many different ways. According to Charpentier (1997), McCleary (1969; 1984) suggested seven types of dasymetric maps, while Robinson et al. (1995) characterize dasymetric maps based on whether they use ancillary information as either limiting variables or related variables. The limiting variable method sets maximum population totals or densities that may be assigned to an area, while the related variables approach explores

the correlation between population distribution and related variables (Robinson et al., 1995). Most often land use and land cover maps derived from aerial photographs or satellite systems are the surrogate variable of choice, but transportation networks, topography, as well as other remote sensing-derived data may be used. McPherson and Brown (2003) and Reibel and Bufalino (2005) recently innovated the use of street databases for dasymetric disaggregation of population.

A specific form of the limiting variable approach is the binary method, the simplest and original form of dasymetry, which is merely the assignment of population to inhabited areas, which are defined by categorical ancillary data. Applied by Fisher and Langford (1996) this method has shown to be less sensitive to land cover classification error (Langford, 2004).

More elaborate dasymetric approaches redistribute census population to different land use types based on estimated population density ratios between these classes. These weights can be estimated using global or regional regression analyses (e.g., Langford et al., 1991), be pre-defined heuristically or empirically (e.g., Eicher and Brewer, 2001), or be determined using empirical sampling techniques (e.g., Mennis, 2003; Mennis and Hultgren, 2006a, 2006b).

Langford et al. (1991) used regression of population density and land use within a dasymetric mapping procedure to redistribute population in the U.K. to a 1-km raster surface. The model overestimated population in urban

areas and underestimated population in suburban and rural regions, showing the difficulty of deriving such models due to the spatial variation in the nature of land use and land cover (LULC) and other variables as they relate to population density. Langford (2006) also refers to the presence of spatial non-stationarity in the relationship between population and LULC.

Using Landsat TM imagery, Harvey (2002a) devised a pixel-level dasymetric method by conducting an ordinary least-squares regression between population density and the digital values of the pixels. In contrast to such complex approaches, Langford (2007) proposed simplicity and convenience by demonstrating the use of raster pixel maps to derive urban masks for dasymetric-based population interpolation.

A comprehensive study by Eicher and Brewer (2001) mapped six socio-economic variables using ancillary land-use data to test the accuracy of five dasymetric and areal interpolation methods: binary methods (in raster and vector), three-class methods (raster and vector), and the limiting variable method (vector). Their analysis concluded that the traditional limiting variable method performed best and that vector implementations had lower error but could not be deemed statistically better than the raster methods, the latter being more easily implemented. However, the adopted 1-km resolution can be considered fairly coarse and scale effects resulting from choice of cell size remain under-researched in dasymetric mapping. Resolution must be fine enough to adequately capture the spatial variation of population distribution.

Mennis (2003) generated raster population surfaces at 100-m resolution for five counties in Pennsylvania, USA, using remote sensing-derived land cover as ancillary data. That author built on Eicher and Brewer's (2001) "grid three-class" method, but instead of using subjectively-determined percentages for disaggregating population to cells, these percentages were determined by empirical sampling densities in three urbanization classes. An area-based weighting scheme addressed relative differences in area among urban land cover classes within a given unit. While Mennis conducted no formal accuracy assessment, Sleeter (2004) applied this approach using the National Land Cover Data set to model population distribution in the San Francisco Bay region (CA, USA) at high spatial resolution (30 m) and obtained high correlation coefficients against census block populations. Sleeter and Wood (2006) applied a similar approach using parcel-level land use and density to map daytime and nighttime population density for a coastal county in Oregon at 10 meters.

Building on efforts by Langford (2006), Mennis later improved his approach by sampling choropleth map zones using centroid, contained, and percent cover rules to associate a zone with an ancillary class (Mennis and Hultgren, 2006a; 2006b).

Wu et al. (2005) remarked that provided that ancillary information reflects the spatial distribution of the variables being mapped, methods using

this information, especially the dasymetric method, tend to be more accurate than those without ancillary information.

Fisher and Langford (1995) classified areal interpolation approaches into cartographic, regression, and surface methods, and upon testing five methods (Fisher and Langford, 1996) showed that cartographic methods provided both the best (dasymetric mapping) and worst (areal weighing) estimates.

#### 2.3.4 Statistical Modeling Methods

The application of statistical modeling to population estimation started in the 1950's, building from theories in urban geography (e.g., Clark, 1951) that related population distribution and morphological factors that often can be derived from remotely sensed data. A feature of the "purely" statistical modeling approach is that, normally, census population data are not used as the input, rather these data are used in the model training phase. Although this approach is "originally designed to estimate intercensal population or population of an area difficult to enumerate, it can also be incorporated into the process of interpolating census population" (Wu et al., 2005, p.1), in what can perhaps be viewed as a hybrid approach.

A plethora of studies have explored correlation between population and numerous variables, such as urban extent (Clark, 1951), multispectral satellite

imagery reflectance (Wu, 2007), satellite-derived vegetation indices (Li and Weng, 2005), impervious surface (Lu and Li, 2006), and others. The predictive power of nighttime lights for large areas especially has been explored and demonstrated by various scientists (Sutton, 1997; Sutton et al., 1997, 2001; Lo, 2001; Pozzi et al., 2003). Harvey (2002b) offers a comprehensive review of methods to estimate population using satellite imagery and their respective strengths and weaknesses.

Methods of “smart interpolation” using multiple variables have also been proposed and implemented by Deichmann and Eklundh in 1991 (Deichmann, 1996), and by Sweitzer and Langaas (1995). Deichmann (1996) summarizes the process in three steps:

1. A surface of weighting factors is created on a regular raster grid for the study area.
2. Initial weights are heuristically adjusted using ancillary data.
3. Total population count for the study area is distributed to cells according to the weights, preserving the total volume.

This approach was adopted for production of the LandScan population database, in which the combination of road proximity, slope, land cover, nighttime lights, and an urban density factor oriented the disaggregation of census data to grid cells in a probabilistic model. LandScan is considered “probably the best representation of rural and urban population density

available” (Sutton et al., 2003) and has been used for many purposes (Dobson, 2002; 2003; 2007; Dobson et al., 2003).

#### 2.4 Scale of Study

Scale is an important factor in determining the usefulness and accuracy of population mapping, and it can be addressed in both the spatial and temporal dimensions. Normally the geographical extent of a study defines the spatial scale, often characterized as local, regional (national) or global. The scale of the study and existence of compatible input datasets has implications for the choice of appropriate resolution (Balk et al., 2006).

In the context of population distribution modeling, the large majority of efforts seem to take place at the fringes of local and global scales, with a scarcity of studies at the national level. This may in part correlate with availability of input datasets, which so often are not made accessible to researchers at the national level or because the national data tend to be prohibitively costly for use by most researchers. In contrast, a number of international institutions and companies have recently begun to have the storage and computing capacity to collect large geographic datasets for use in global studies, even if their characteristics are neither homogenous nor ideal. Therefore, most experimental research efforts take place at the more controlled local or regional scales, where knowledge of the study area may

assist in dealing with data and interpreting results. Some examples are Langford et al. (1991), Goodchild et al. (1993), Fisher and Langford (1996), Eicher and Brewer (2001), Mennis and Hultgren (2006a; 2006b), and Langford (2006; 2007).

Regarding global scale models, the 1990's can be considered the age of global gridded population datasets, with three new undertakings adding to previous efforts by the US Bureau of Census (Leddy, 1994): the Global Demography Project / Gridded Population of the World (GPW) (Tobler et al, 1995), the HYDE dataset (Goldewijk and Battjes, 1997), and the LandScan Global Population Database (Dobson, 2000; Dobson et al, 2000). Their main characteristics are summarized in Table 2.

Table 2. Characteristics of global gridded databases of population density

<b>Project</b>	<b>Resolution (cell size)</b>	<b>Temporal coverage</b>	<b>Spatial coverage</b>
<i>LandScan Global Population Project</i>	30"	1998 – 2005	World
<i>Gridded Population of the World, v3</i>	2.5'	1990, 1995, 2000	World (72 N to 57 S)
<i>HYDE Database, v. 2.0</i>	30'	1700 – 1995, yearly	World
<i>Global Population Database</i>	Variable: 20' X 30' and 5' X 7.5'	Variable, starting in 1965	Selected countries, 150 in 1994

In fact, some initial efforts have evolved into quasi “operational” programs featuring subsequent versions whose periodical release mostly



reflects improvements in spatial and temporal resolution of the source data. Notable examples are GPW, with three versions released in 1995, 2000 and 2004 (Deichmann et al., 2001; Balk et al., 2006), and LandScan, with releases for 1998, 2000, 2001, 2002, 2003, 2004, and 2005 (ORNL, 2007).

### 3. METHODOLOGY

This section presents the study area, describes the main procedures and software used, the collection and organization of variables, the modeling of population distribution, and procedures used to verify and validate the results.

#### 3.1 Study Area

The official administrative limits of the municipalities (*concelhos*) of Cascais and Oeiras in 2001 constitute the study area for this research. This area was selected for several reasons: a) its characteristics, namely with regard to urban and suburban character, and strong economic activity, b) the availability and access to input data, and c) personal familiarity with the area, which facilitates data verification and field work.

Cascais and Oeiras are two of the eighteen municipalities that comprise the Lisbon Metropolitan Area (LMA) in Portugal. This is the main metropolitan area in the country with a resident population of 2,661,850 (INE, 2001) encompassing a surface area of 2,963 square kilometers, representing 26% of the Portuguese population and 3.2% of the national area. These two adjacent municipalities are located at the mouth of the Tagus river and along

the Atlantic coast immediately to the west of the city of Lisbon. They border the municipalities of Amadora and Sintra to the north (see Figure 1).

The topography of the study area varies from flat to rolling, but is carved with local valleys created by small streams flowing north-south. Most of the land is occupied by artificial surfaces, with some areas of agriculture and natural vegetation remaining. The western part of Cascais is classified as a Natural Park. Regarding the settlement pattern, there is both concentrated and dispersed settlement, with the latter being more prevalent in Cascais due to the rapid and unplanned urbanization taking place in the 1970s. As is typical of Portugal (see Freire and Caetano, 2005), urbanization is strongest along the coast and is expanding to the interior.

These municipalities have a number of characteristics that make them good candidates for this study. Eighteen percent of the resident population in the study area commutes to Lisbon for work or school (INE, 2003a). Despite the attracting effect of Lisbon proper for employment, these two municipalities also display intense economic activity within their boundaries. Most activities belong to the tertiary sector, but in Cascais the secondary is still an important economic factor. Oeiras recently created technological and office parks, while Cascais maintains some industrial zones. Until the 1960s this coastal area was mostly sought for leisure, and tourism is also a significant activity.

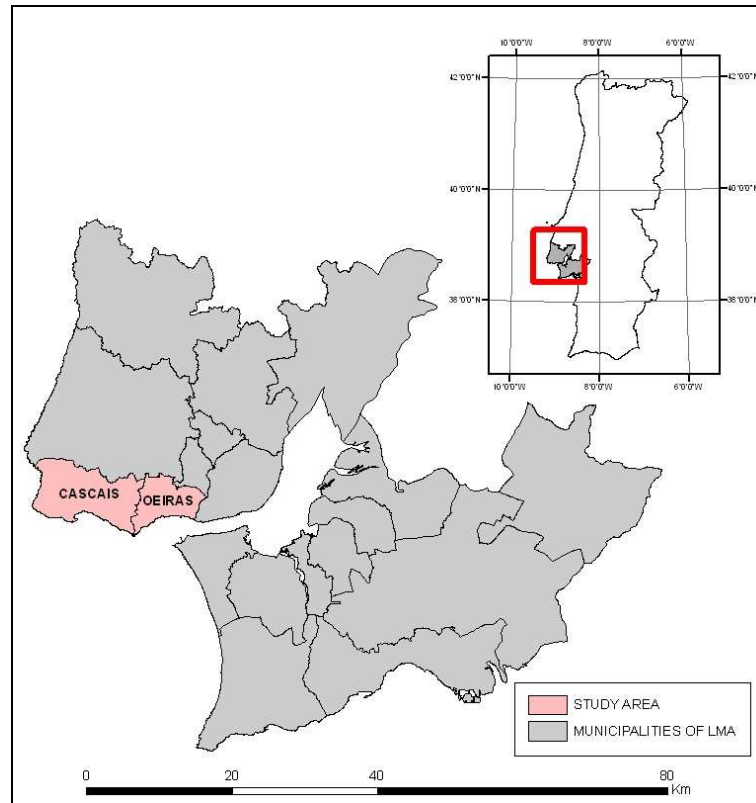


Figure 1: Location of the study area in Portugal and in the Lisbon Metro Area.

Cascais and Oeiras occupy 97 and 46 km<sup>2</sup>, respectively, and have a combined population of 332,811; this results in an average population density of 2332 inhabitants/km<sup>2</sup>, well above the national average density of 112 inhabitants/km<sup>2</sup>. However, population density varies widely throughout the study area, from high density in multi-story residential apartments to low-density in rural areas.

The resident population of the study area, the LMA and Portugal as a whole and their evolution in the last three national censuses (1981, 1991, and 2001) are reported in Table 3.

Values show that population in the study area has been increasing well above the national rate, especially from 1991 to 2001, and more so in Cascais, which also grew more rapidly than the metro area in both periods.

Table 3. Population and growth rates in census years, 1981-2001.

	<b>1981</b>	<b>1991</b>	<b>Change 81-91 (%)</b>	<b>2001</b>	<b>Change 91-2001 (%)</b>
Cascais	141,498	153,294	8.3	170,683	11.3
Oeiras	149,328	151,342	1.3	162,128	7.1
LMA	2,502,044	2,540,276	1.5	2,682,687	5.6
Portugal	9,833,014	9,867,147	0.3	10,356,117	5

*Source: INE, 2003a; 2003b*

### 3.2 Description of Software

There were several steps involved in attaining the goals of this study, namely collection, organization, and processing of variables, modeling, and quality assessment of population grids. Software used for pre-processing of data, modeling, analysis and presentation of results included the following:

1. Microsoft Access – a database management system used for creating, managing and querying large data sets. This software was used to query and retrieve census population data.
2. Microsoft Excel – a spreadsheet used for organizing data in tables and analyzing small data sets. This software was used to

compile and verify data, to create tables for importing to GIS software, and to analyze and produce statistics.

3. ESRI ArcView – a desktop mapping and Geographic Information Systems (GIS) application used for collecting, managing, integrating, analyzing and presenting geospatial data sets. This software was instrumental for inspecting and improving input data for modeling.
4. ESRI ArcGIS – a more powerful mapping and Geographic Information Systems (GIS) application used for collecting, managing, integrating, analyzing and presenting geospatial data sets. This software was essential for geocoding addresses of workers and students for daytime population modeling, for other analyses, and to produce maps to present the results.
5. Google Earth (<http://earth.google.com/earth.html>) – a downloadable application used to visualize geospatial information using web-based satellite imagery as a base map. This application was useful for comparing and improving input data.

### 3.3 Collection and Organization of Variables

In geographic modeling, choice of input variables and their characteristics should be appropriate to the objectives and mutually compatible in terms of scale, resolution and reference date. In particular, assuring temporal consistency of data sets is usually challenging and often impossible in the context of population mapping, even at local or regional scales: for instance, Eicher and Brewer (2001) redistributed 1990 census data using a 1970 land use and land cover (LULC) map, while Mennis (2003) based disaggregation of 1990 population counts on 1993/1996 urban land cover data.

In this study, the spatial detail of census zones whose counts are to be disaggregated should be met in scale, resolution, and accuracy by ancillary data sets used for disaggregation. Input variables used for modeling include both physiographic and statistical data. In the first group are census tracts, street centerlines and land use and land cover (LULC), while the second includes census counts, data on workplaces and employment, and commuting statistics for the study area. These data were obtained from various sources and in different formats which are listed in Table 4.

In general, temporal consistency among data sets was very high, with the exception of street centerlines whose reference date was three years subsequent to the model target date (2001). For this reason and due to the

importance of this data set in the model, it was decided to modify it in order to better represent reality at the target date. The one-year lag in LULC data was considered acceptable. A detailed description of, and modifications applied to, each variable listed follows.

Table 4. Main input datasets used for modeling nighttime and daytime population.

<b>Data set</b>	<b>Source</b>	<b>Date</b>	<b>Data type</b>
Street centerlines	GeoPoint	2004	Vector polyline
LULC (COS90; CLC2000)	IGP; IA	1990; 2000	Vector polygon
Census block groups	INE	2001	Vector polygon
Census statistics	INE	2001	Database (MS Access)
Workplaces and employment	DGEEP	2001	Table
Commuting statistics	INE	2001	Table (O/D matrix)

### 3.3.1 Street Centerlines

Street centerlines represent streets through a single line digitized along their center, and are most useful for address geocoding (or address matching). These data are characterized by their high spatial detail, positional accuracy, and ease of updating. Furthermore, in the framework of this project they provide a convenient support for modeling population by allowing the same spatial basis to be used for both nighttime and daytime distributions. The fact that street centerlines are provided for spatial reference by the US Bureau of Census (i.e., the TIGER/Line files) confirms importance of its relation to socioeconomic variables. Reibel and Bufalino (2005) used street



grids without distinction of classes of streets or roads and still obtained better results than areal weighting alone. Sleeter and Wood (2006) recognize that use of a roads layer in their mapping of daytime population would improve land use delineations and results.

Street centerlines for each of the two municipalities for 2004 were obtained (in MapInfo .MIF format) from GeoPoint ([www.geopoint.pt](http://www.geopoint.pt)), a private vendor which compiles this information from several sources: municipalities, other companies, and the postal service (see Figure 2). Unfortunately a record of previous versions was not kept and it was impossible to obtain such data for 2001. These data were very detailed and complete, both geometrically and regarding attributes used for geocoding addresses.

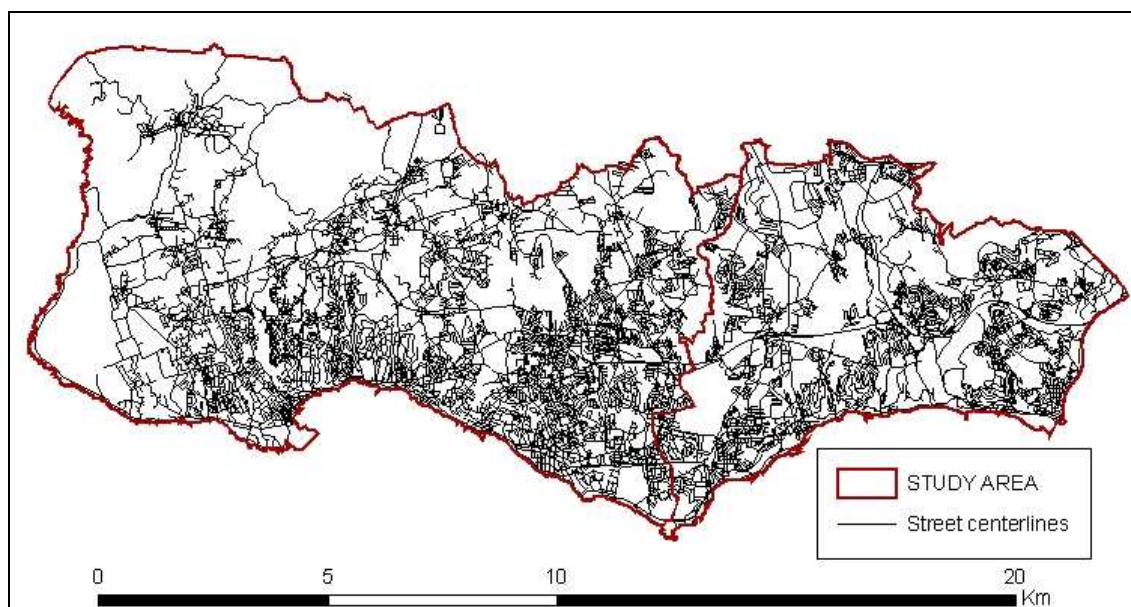


Figure 2: Street centerlines obtained for the study area.

### 3.3.2 Land Use and Land Cover

Initial information on land use and land cover was obtained from two data sets in the public domain available in vector format for mainland Portugal: the Land Cover Map of 1990 (COS'90), with a minimum mapping unit (MMU) of 1 ha and a complex classification system that allows more than 900 classes (see Figure 3); and the CORINE Land Cover 2000 (CLC2000) map (Figure 4), with a 25-ha MMU and a hierarchical nomenclature with 44 classes in the most detailed version (level 3) (Bossard et al., 2000). The CLC nomenclature is presented in Table 5. Despite their designations, both data sets aim at mapping both land use and land cover classes. Since imagery only captures land cover, these maps were produced using visual interpretation aided by ancillary data and field work to infer land use, which is a very challenging process.

CLC is an operational European program and its data were already used as ancillary information to disaggregate population counts in the European Union (Gallego and Peedell, 2001). In that experiment, population density was mapped to 100-m cells (1 ha), but validation showed that CLC alone was insufficient to accurately represent population in the study area.

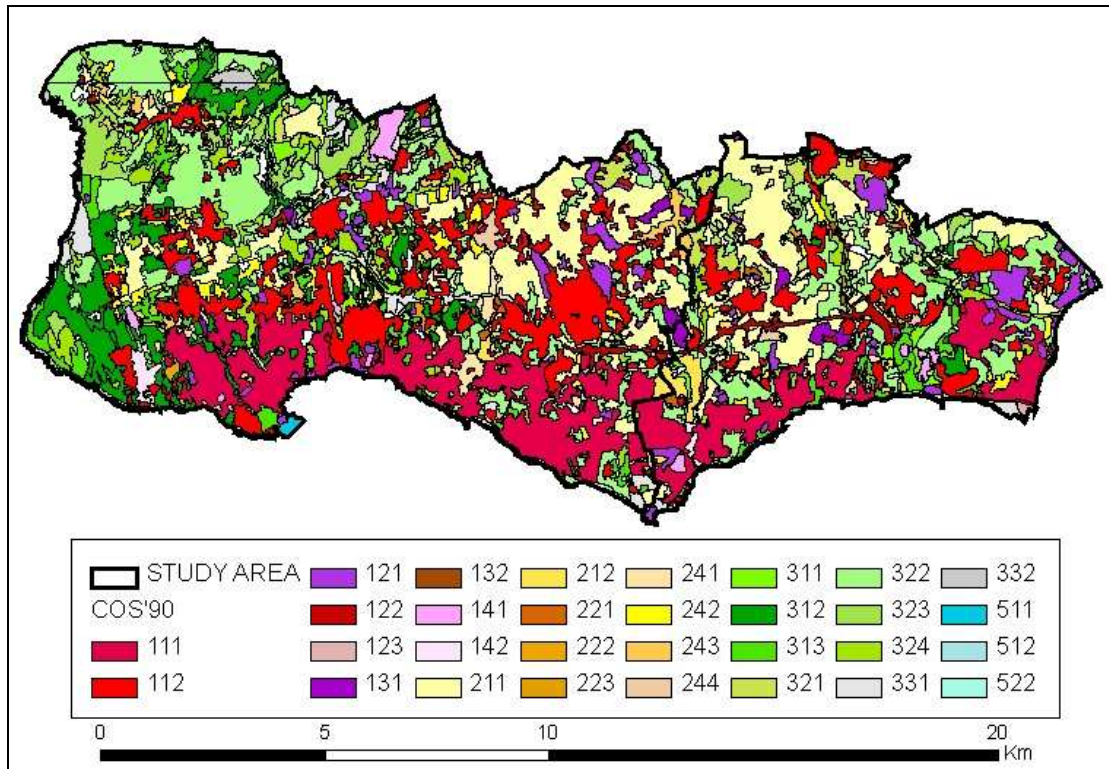


Figure 3: The Land Cover Map of 1990 (COS'90) for the study area, converted to CLC nomenclature. See Table 5 for an explanation of the land cover classes

For modeling population, accurate mapping of urban residential land use is important (Langford, 2004), and an accurate LULC data at appropriate resolution can adequately represent population distribution, potentially dispensing with the need for other variables. However, no formal accuracy assessment was conducted for COS'90, while validation of CLC2000 in Portugal has shown a thematic agreement of 83% (Caetano et al., 2006), despite the fact that the evaluation was not designed as a strict accuracy assessment.

Another issue impacting the use of these data for modeling has to do with size of MMU and the classification system used, and the fact that LULC classes are mutually exclusive. This creates problems in areas where land uses are mixed (either horizontally or vertically) but had to be classified under only one “pure” class (e.g., residential-commercial areas classified as residential). Perhaps this problem could be mitigated by including additional mixed classes in the LULC nomenclatures or using a main class and a secondary class to characterize mixed polygons.

Upon careful inspection, important errors were detected in these data sets within the study area. Because the nighttime population surface is so dependent on correct identification of residential areas, LULC data has to be the most accurate possible, and so it was decided to correct them prior to modeling.

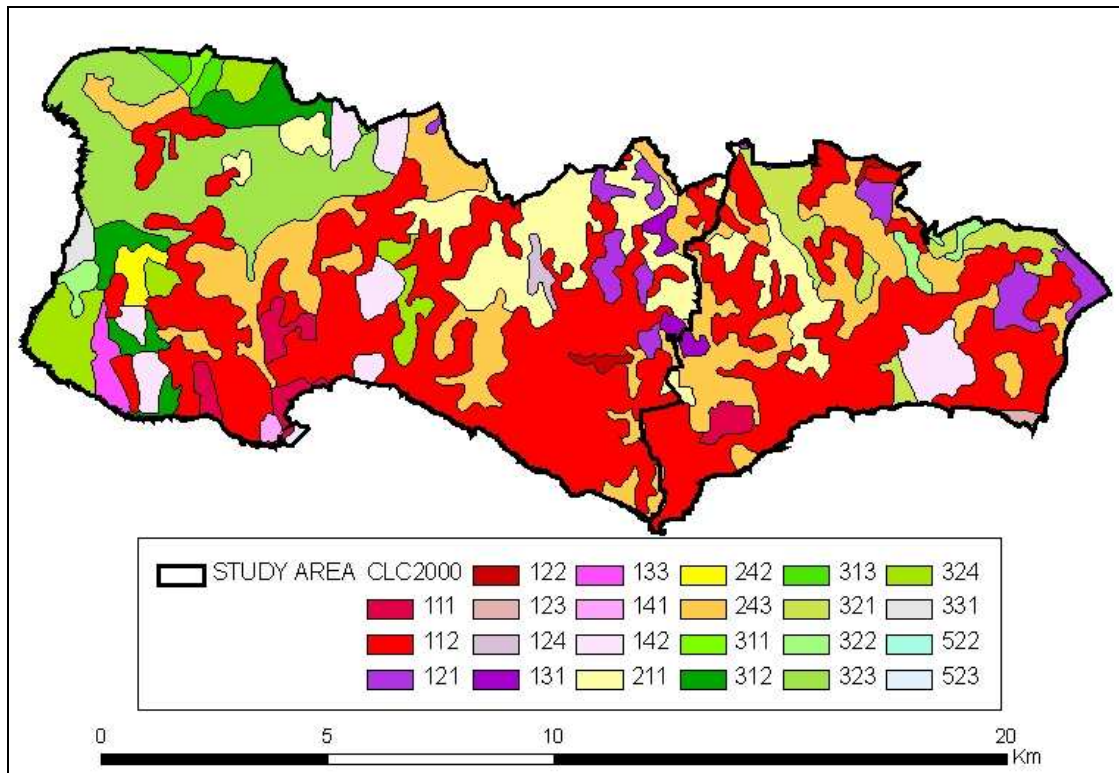


Figure 4: The CORINE Land Cover 2000 (CLC2000) map for the study area. See Table 5 for an explanation of the land cover classes

Table 5. Standard CORINE Land Cover nomenclature.

Level 1	Level 2	Level 3
1. Artificial surfaces	1.1 Urban fabric	1.1.1 Continuous urban fabric 1.1.2 Discontinuous urban fabric
	1.2 Industrial, commercial and transport units	1.2.1 Industrial or commercial units 1.2.2 Road and rail networks and associated land 1.2.3 Sea ports 1.2.4 Airports
	1.3 Mines, dumps and construction sites	1.3.1 Mineral extraction sites 1.3.2 Dump 1.3.3 Construction sites

	1.4 Artificial non-agricultural vegetated areas	1.4.1 Green urban areas 1.4.2 Sport and leisure facilities
2.Agricultural areas	2.1 Arable land	2.1.1 Non-irrigated arable land 2.1.2 Permanently irrigated land 2.1.3 Rice fields
	2.2 Permanent crops	2.2.1 Vineyards 2.2.2 Fruit trees and berries plantations 2.2.3 Olive groves
	2.3 Pastures	2.3.1 Pastures
	2.4 Heterogeneous agricultural areas	2.4.1 Annual crops associated with permanent crops 2.4.2 Complex cultivation patterns 2.4.3 Land principally occupied by agriculture with significant areas of natural vegetation 2.4.4 Agro-forestries
3. Forest and semi-natural areas	3.1 Forests	3.1.1 Broad leafed forest 3.1.2 Coniferous forests 3.1.3 Mixed forest
	3.2 Scrub and/or herbaceous vegetation associations	3.2.1 Natural grassland 3.2.2 Moors and heathlands 3.2.3 Sclerophyllous vegetation 3.2.4 Transitional woodland-scrub
	3.3 Open spaces with little or no vegetation	3.3.1 Beaches, dunes, sand 3.3.2 Bare rocks 3.3.3 Sparsely vegetated areas 3.3.4 Burnt areas 3.3.5 Glaciers and permanent snowfields
4. Wetlands	4.1 Inland wetlands	4.1.1 Inland marshes 4.1.2 Peat bogs
	4.2 Coastal wetlands	4.2.1 Salt marshes 4.2.2 Salines 4.2.3 Intertidal flats
5. Water bodies	5.1 Continental waters	5.1.1 Stream courses 5.1.2 Water bodies

	5.2 Marine waters	5.2.1 Coastal lagoons 5.2.2 Estuaries 5.2.3 Sea and ocean
--	-------------------	---

### 3.3.3 2001 Census Data

The National Statistical Institute of Portugal (INE) conducts decadal censuses of population, dividing the basic administrative level (*Freguesia*, i.e., commune) into two-level statistical areas called *subsecção estatística*, and *secção estatística*. These areas are defined by polygons in a geographical base for referencing with attributes in a database. In this study, counts of resident population (i.e., *de jure*) from the 2001 census were used at the level of *secção estatística* (i.e., similar to block group in the US) for modeling, with block-level data (i.e., *subsecção estatística*) used for validating results. One *secção estatística* unit corresponds to a continuous area of a single *Freguesia* (i.e., commune) with approximately 300 dwellings used for residence. Counts of present (i.e., *de facto*) population are also available, but these fail to represent population in workplaces or schools in a normal weekday.

In the study area there are 529 block groups, with all of them having some resident population. However, the population density of block groups varies significantly from urban to more rural parts of the region, and some polygons misrepresent actual population presence due to their large size. The



mean size of block groups in the study area is 27 ha, with a standard deviation of 72.4 ha.

Population density by block group in the study area is presented in Figure 5, classified by quantiles.

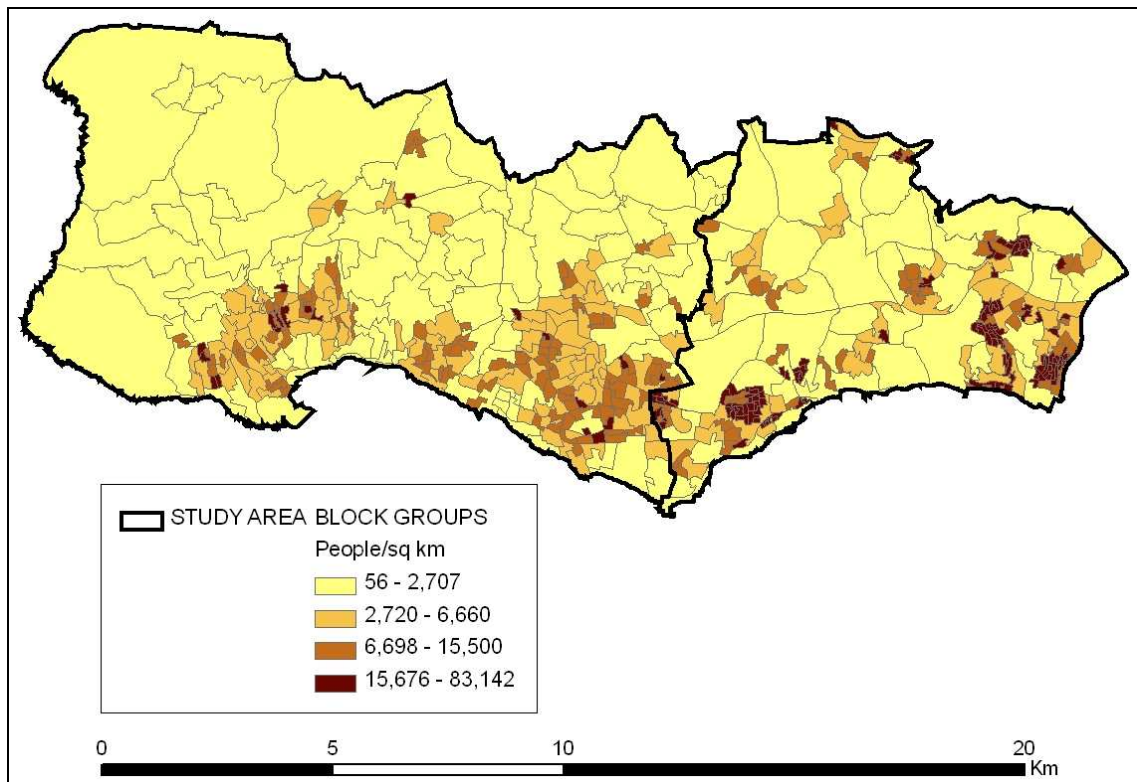


Figure 5: Density of resident population in the study area, by *secção* (block group).

The INE produces inter-censal estimates for each year at the municipal level, but due to the aggregated level and low confidence these data are unfit for local-level modeling of population.



#### 3.3.4 Workplaces and Employment

The main data set on workforce was acquired from the Directorate-General of Studies, Statistics, and Planning (DGEEP) of the Ministry of Labor (DGEEP, 2001). Data were obtained as a table listing for Cascais and Oeiras the name of workplaces (firms), address of facilities, business code and type, and number of employees in 2001. Companies involved in temporary or displaced activities (such as catering, construction) are also included, but specific addresses are not provided. Because DGEEP collects only data on private firms, it was necessary to complement this information with data on public workplaces.

A shapefile locating schools in the Metro Area and a database listing respective students and workforce was obtained from the Lisbon Metro Area services.

#### 3.3.5 Commuting Statistics

Data on number of daily commuters for reasons of work or school in 2001 were obtained from an official mobility study by INE (2003a) for the municipalities integrating the Metropolitan Areas of Lisbon and Porto. This is the most recent and most detailed disaggregated data yet available on the subject. Figures were available in paper as an origin/destination (O/D) matrix

recording the number of people leaving their homes for the cited reasons and leaving, entering, and remaining in each municipality. This study provided the municipal reference totals for daytime population distribution and/or control: total workforce, total daytime residential population, and total daytime population.

Geographical data were transformed into a convenient format for modeling in GIS: vector data sets were converted to shapefile format (ESRI) and to the national projection system used for large-scale mapping (scale 1: 25,000). The national projection system uses the Hayford-Gauss Militar projection and the Lisbon Datum. Although this projection is conformal and not equal-area, distortions in the study area are negligible in the context of the study.

### 3.4 Modeling the Population Distribution

In the present context, modeling population distribution at a specific spatio-temporal scale entails two basic challenges: a) obtaining or deriving reliable population counts at an appropriate spatial and temporal level for disaggregation, and b) applying a method to interpolate those counts in a realistic fashion. In the current approach, two types of information are combined: a) statistical and census data and b) physiographic data. The most recent statistical and census data (2001) provide the population counts for

each daily period, while physiographic data sets define the spatial units (i.e., grid cells) used to disaggregate those counts. This general approach was successfully implemented by Los Alamos National Laboratory (New Mexico, USA) to map daytime and nighttime population distributions in the US at 250-m resolution (McPherson and Brown, 2003; McPherson et al., 2004; 2006), and is adapted and applied to Portugal.

This method allows the daily temporal component of population to be considered for estimating their nighttime and daytime distributions and was implemented within a Geographic Information System, ArcGIS 9.1 (ESRI). GIS offers the necessary tools and flexibility to implement raster or vector-based dasymetric methods, and was used to verify, correct and integrate geographic data sets, for modeling, analysis, and mapping the results for presentation.

Because some data sets are made available on a municipal basis, Cascais and Oeiras were modeled separately. For the area of each municipality, 25-m raster grids were produced representing densities of nighttime (residential) population, daytime residential population, daytime worker and student population, total daytime population, and ambient population in 2001.

The raster format offers several advantages: it constitutes a regular tessellation of space as square cells, which is convenient for GIS analysis and modeling. The grid cells are flexible reference units appropriate for

representing numerical data and they facilitate relating data from incompatible areal units using GIS zonal functions (Martin and Bracken, 1991). Furthermore, most environmental and simulation models (e.g., diffusion of pollutants) in which population exposure is analyzed also adopt the raster data model.

Regarding spatial resolution, its choice should depend on the goals and scale of the study (Balk et al., 2006) and on the existence of compatible input datasets. Considering these principles it was decided that 25 meters was the highest resolution possible to model population with confidence in this project. Modeling at the highest spatial resolution possible seems to make more sense at a time when storage and processing capacity become less of an issue: fine raster data can easily be aggregated to coarser while coarser results cannot be adequately refined without modeling from source data again.

A cell size of 25 meters was selected for the following reasons:

1. it suits the objective of supporting local-level analysis;
2. it approximates dimensions of an average single-family lot (half-block width), and allows modeling at the scale of individual buildings without compromising statistical privacy, required in Portugal;
3. it is a sub multiple of 50 and 100 meters, which are potential cell sizes for public dissemination of the data;

4. it is smaller than the dasymetric zones used in this study.

Following the suggestion by Eicher and Brewer (2001), a limited sensitivity analysis on the effect of grid cell size on accuracy for the adopted method was conducted for the municipality of Oeiras. Nighttime population distribution was mapped at resolutions of 12.5 m, 25 m, and 50 m, all of these cell sizes being fine enough for local-level applications in accordance with the project's goals. Correlation analysis results comparing aggregated population counts from each final surface to official census blocks are presented in Table 6.

Table 6. Results of correlation analysis of cell size in Oeiras.

Cell size	No. of cells	Correlation	R <sup>2</sup>
12.5 m	33,119	0.80	0.64
25 m	15,342	0.79	0.62
50 m	6,255	0.75	0.56

The analysis shows that while resolution consecutively increases four-fold, there is only a small improvement in accuracy with 25-m cells compared with 50 m, and this improvement becomes marginal with an increase in resolution from 25 m to 12.5 m. Therefore this study selected the 25 m grid resolution.

The main phases involved in creating the population surfaces included pre-processing of data sets, modeling the distributions, and verification and validation of results. An overview of tasks involved in the modeling process is presented in Figure 6.

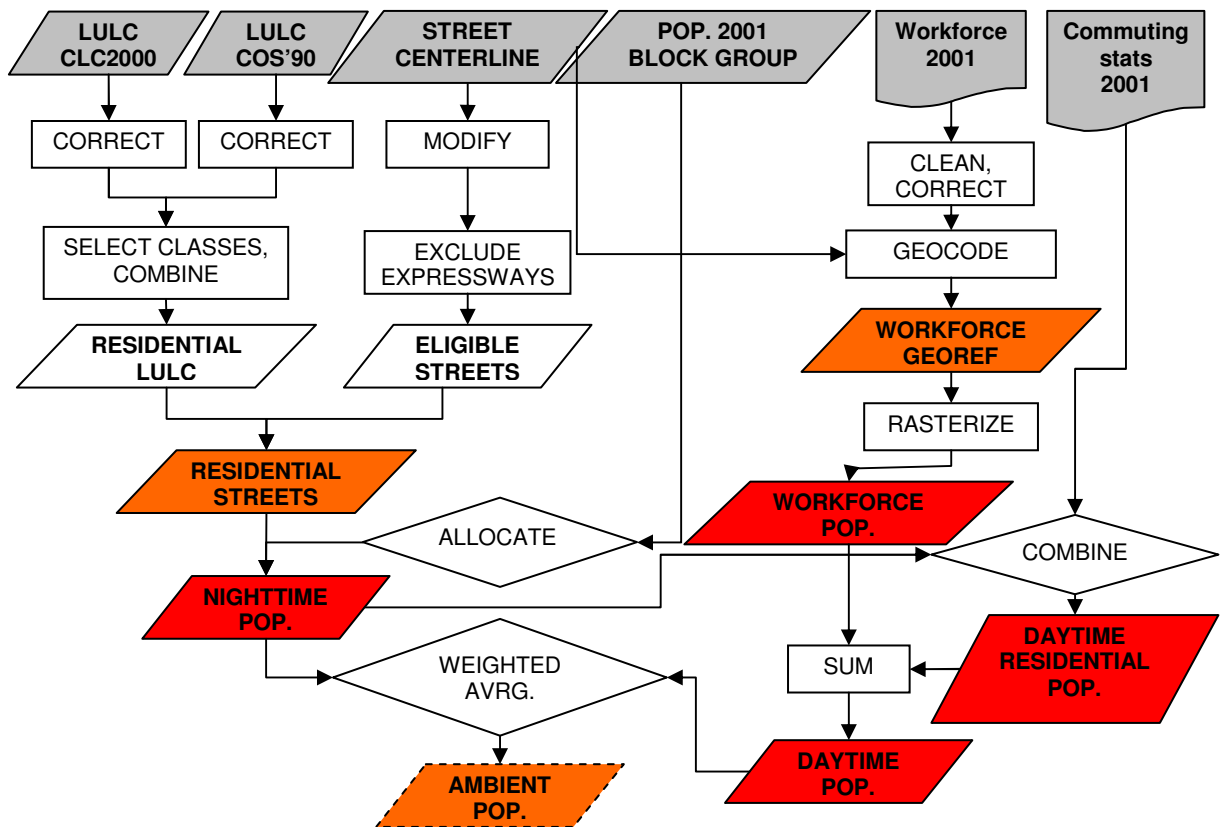


Figure 6: Flowchart of main tasks involved in modeling the spatial and temporal distribution of population.

### 3.4.1 Pre-processing of Data Sets

Due to the characteristics of the data and the model requirements, pre-processing of some data sets was necessary:

1. The five COS'90 individual worksheets (map tiles) covering the study area were downloaded and merged together. This shapefile was then improved using ArcView based on orthophotos for 1995, municipal paper maps, and personal knowledge of the area. The main problems detected and corrected involved misclassification of industrial or commercial units as residential areas. With correction, the number of polygons in this COS'90 shapefile changed from 1162 to 1259. Figure 7 illustrates corrections to COS'90 made for a coastal area in Cascais.

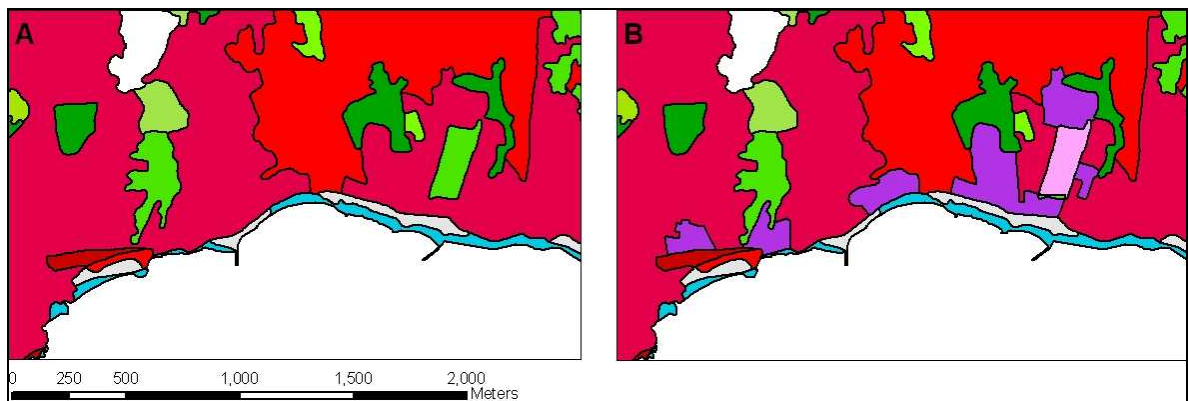


Figure 7: Sample of original (A) and corrected (B) COS'90 LULC map.

The example illustrates the identification and separation of industrial or commercial units (class 121) misclassified as urban fabric (classes 111 and

112). Correct delineation of these classes is especially important for model results.

Finally, polygon boundaries were dissolved based on the corresponding CLC codes, using an existing conversion table. The steps taken to pre-process the COS'90 are represented in Figure 8.

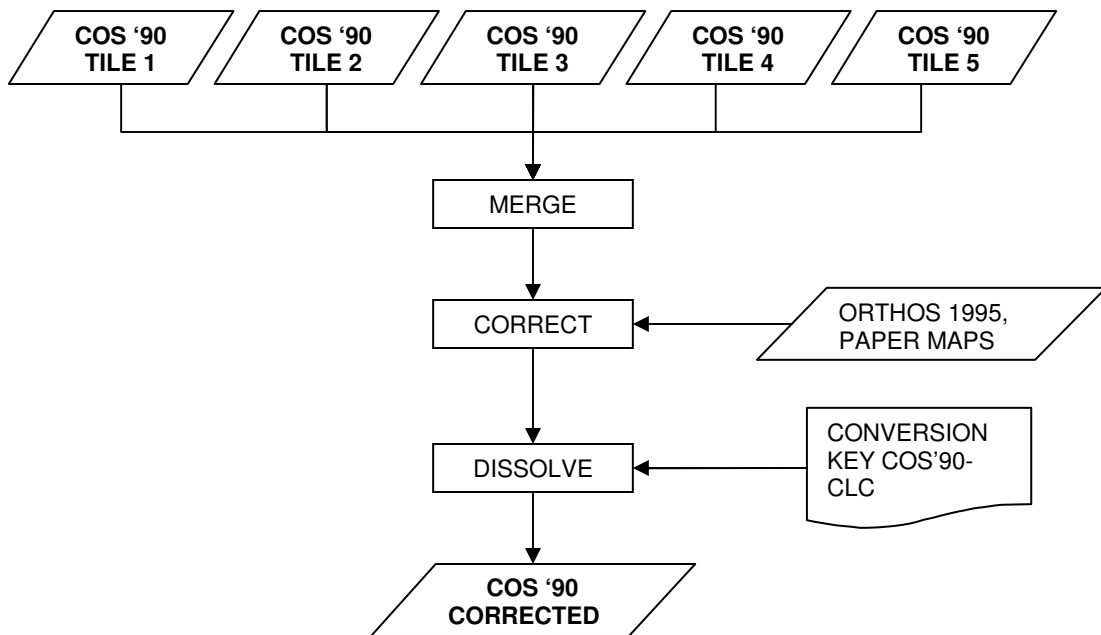


Figure 8: Flowchart of pre-processing of COS'90 LULC map.

2. The CLC2000 map was also inspected and corrected in the study area, using orthophotos for 1995, municipal paper maps, and personal knowledge of the zone. Both thematic and geometric corrections were made to the level 3 CLC2000 nomenclature, with the number of polygons increasing from 107 to 113 in the study area. Most problems involved the



misclassification of industrial or commercial land use as residential. Figure 9 illustrates corrections to CLC2000 made for an area in Oeiras.

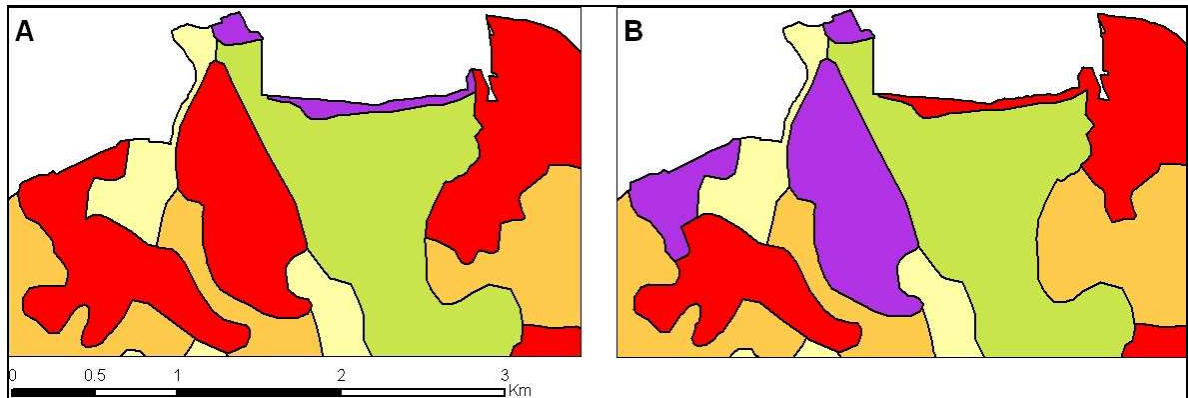


Figure 9: Sample of original (A) and corrected (B) CLC2000 LULC map.

3. Street centerlines for each of the two municipalities for 2004 were converted from MapInfo .MIF format to ESRI's shapefile (.shp) format. Due to the importance of streets in the model and the rate of change in the area, it was decided to modify these data to better represent the street network existing in 2001. This editing was accomplished in ArcView with the aid of ancillary data (paper maps, satellite imagery, Google Earth). Figure 10 shows the street network obtained for 2004 and edited to represent 2001. Red lines correspond to streets built or modified between 2001 and 2004 and which were corrected or removed from the street shapefile used for modeling.

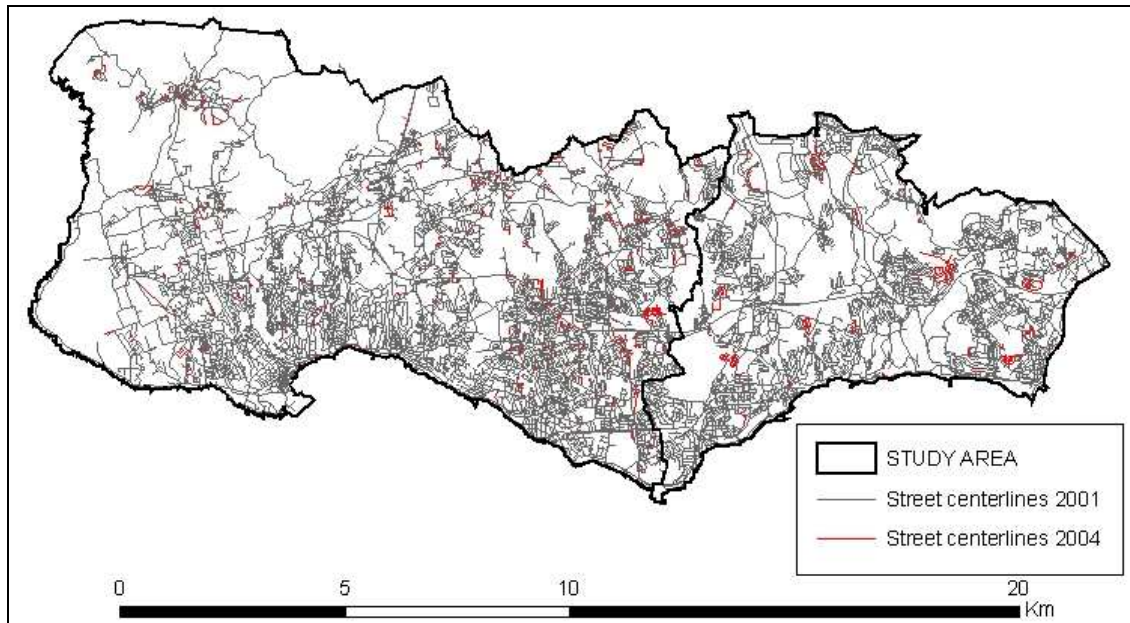


Figure 10: Street network in 2004 and in 2001.

4. Residential population counts for *secção estatística* (i.e., block group) in the study area were selected from the Census database and retrieved as a table, to join with corresponding polygons in the GIS and create the census shapefile.

5. Data on workforce from DGEEP used for placement of workers at facilities were exhaustively verified for errors and improved with the aid of on-line resources such as the Yellow Pages (<http://pai.pt/>) and the companies' own websites. Main problems detected and corrected included insufficient address information and misplacement of workplaces among municipalities. Overall, it was found that 78 of the 4129 workplaces listed in the DGEEP database were not located in the study area and were removed.

### 3.5 Nighttime Population Distribution

Nighttime population distribution is modeled by combining a dasymetric mapping technique with an areal interpolation method to disaggregate residential population to residential streets. A grid binary method is employed to model this distribution: total resident population is evenly distributed to selected grid cells in each *seção estatística* (block group). The binary method represents a specialized form of the limiting variable method described by McCleary (1969), in the sense that all land uses deemed uninhabited are limited to zero population (Eicher and Brewer, 2001).

This approach can also be considered a method of spatial interpolation, i.e., transferring data from source zones (census block groups) to non-nested or incompatible target zones (residential grid cells). The method was selected for simplicity of implementation and because it was one of the most accurate in previous testing (Eicher and Brewer, 2001). Furthermore, this method is well suited for quantitative data contained in polygons, such as population, and allows for volume preservation. Volume preservation is important to this study and a basic requirement for accurate representation of the variable at multiple scales. For raster data this implies that the sum of all grid cell values in each source zone equals its initial total. For the pycnophylactic condition to be met for the study area, all block groups

having population (source zones) must contain at least one residential grid cell (i.e., one dasymetric zone) to receive population.

The main features of the nighttime modeling approach are summarized in Table 7.

Table 7. Summary characteristics of population modeling approach

<b>Technique</b>	<b>Interpolation method</b>	<b>Source zone</b>	<b>Target zone</b>	<b>Resolution</b>
Grid binary dasymetric mapping	Areal weighting	Census block group	Residential street cells	25 m

This grid binary technique requires that only residential streets be considered to receive population. However, street centerlines obtained did not include any attribute discriminating the type of land use. Therefore, land use and land cover maps were used as ancillary information to help identify streets of a residential nature and use. Streets would be considered residential whenever they overlapped residential LULC polygons, with the exception of thoroughfares which do not allow direct access to residences. So, first all expressways and limited access highways were excluded from the street dataset, and a shapefile of eligible streets was thus obtained.

The data set of residential LULC polygons for the study area was created from the previously corrected COS'90 and CLC2000 maps. The rationale was to rely primarily on land use information from the more detailed but somewhat outdated COS'90 map and update it with new urban residential

areas from CLC2000. The CLC level 2 class 11 (Urban fabric) reasonably approximates residential land use, and therefore all corresponding level 3 polygons classed 111 (Continuous urban fabric) and 112 (Discontinuous urban fabric) were selected in CLC2000 and COS'90 and combined in a polygon union operation. However, since the CLC2000 map fails to capture smaller land use types that are mostly immutable in a period of ten years, it was decided to exclude areas classified as such in 1990 from consideration as residential in 2000. Upon inspection of the study area, polygons classified as 121 (Industrial or commercial units), 123 (Sea ports), 124 (Airports), 141 (Green urban areas), and 142 (Sport and leisure facilities) in COS'90 were deemed immutable and selected so their area could be erased from the residential shapefile. This process was implemented in ArcGIS and is outlined in Figure 11.

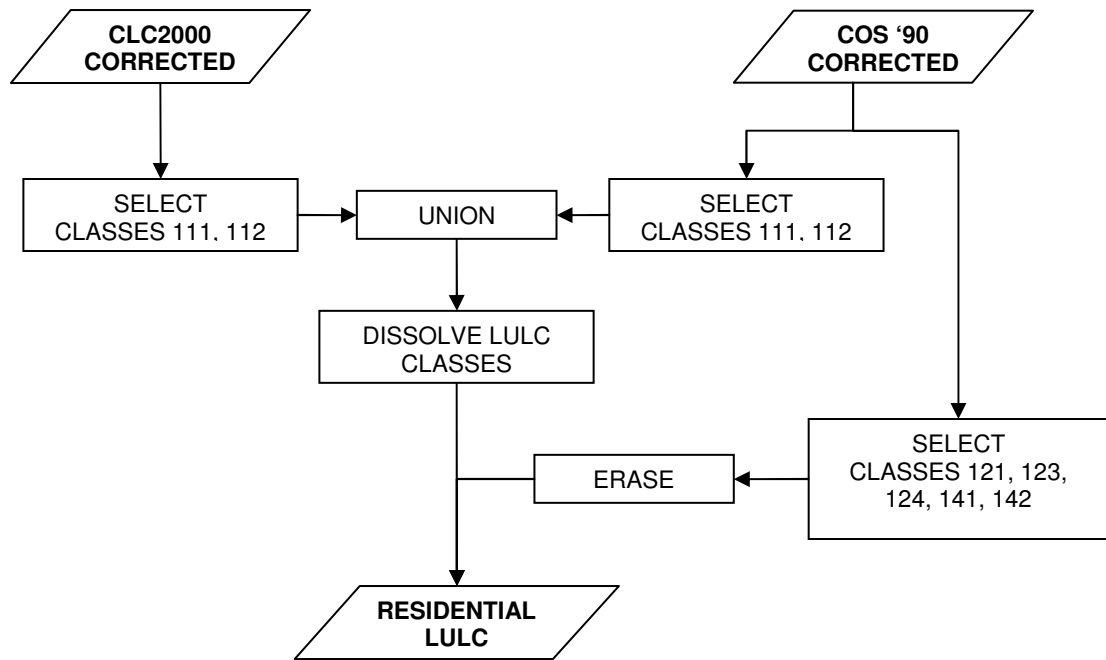


Figure 11: Flowchart of GIS procedure for definition of residential land use polygons

Next, the shapefile of eligible streets was clipped with the residential LULC polygons to derive the shapefile of residential streets to which population could be allocated. Figure 12 shows final residential streets for Oeiras in comparison with all initial street centerlines, as well as residential LULC used in their definition.

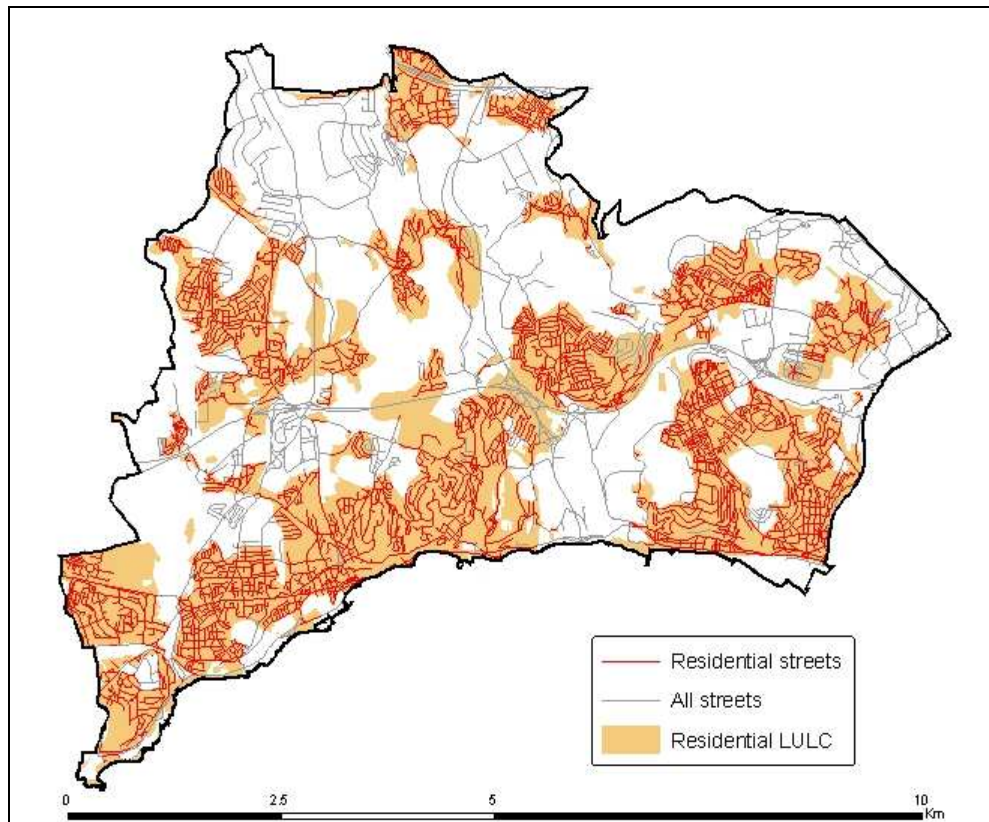


Figure 12: Complete street network, residential polygons and residential streets in Oeiras.

The availability of two level-3 LULC classes representing Urban fabric (codes 111 and 112) could suggest adoption of the limiting variable method, in alternative to the simple binary approach used. However, that solution was not adopted for the following reasons:

1. Defining two residential street density classes based on LULC and weighting them accordingly would most probably be redundant due to the fact that the street network is already denser in areas with land use class 111 than for 112; i.e., varying road densities within source zones should already

account for different population distribution characteristics represented by LULC classes.

2. Distinction between classes 111 and 112 in mapping land use and land cover does not aim at capturing different densities of settlement, but instead correlates with proportion of impervious surface; therefore class 111 may include large parking lots where population density is in reality lower than that of single-family homes classified as 112.

3. It would require setting of threshold densities (weights) for each class. There is no evidence (empirical or other) available to do so for the study area.

Therefore, instead a binary distinction of residential vs. non-residential land use was preferred, with population being distributed evenly within dasymetric zones.

The distribution of population from census block groups to residential streets was calculated according to McPherson and Brown (2003). This was carried out in ArcGIS with the Spatial Analyst extension, using map algebra. First, vector residential streets were converted to raster format. In ArcMap, each residential street shapefile was converted to 25-m grids using municipal limits as a mask, and cells in each block group polygon were counted and written to a raster. This raster was then used to generate a new grid of residential location coefficients necessary to interpolate the census polygon



data. The residential location coefficient represents the proportion of the total residential population in a block group that is allocated to each grid cell and is calculated as:

$$P_{ri} = 1/R_{bg} \quad (1)$$

where  $P_{ri}$  is the residential location coefficient for each cell  $i$ , and  $R_{bg}$  is the number of residential grid cells per block group.

The grid of nighttime residential population was computed according to equation 2:

$$NRP_i = RP_{bg} * P_{ri} \quad (2)$$

where  $NRP_i$  is nighttime residential population in each grid cell  $i$ ,  $RP_{bg}$  the residential population in each block group, and  $P_{ri}$  is the residential location coefficient for each cell.

### 3.6. Daytime Population Distribution

The daytime population distribution is derived from two components (as illustrated in Figure 6): a) the daytime population in their places of work or

study – the workforce population surface, and b) the population that remains home during the day - the daytime residential population density.

For the latter, in the absence of other information, it is assumed that all individuals who according to INE (2003a) do not commute to work or school remain in their residences in the daytime period. This means that this study is not including the potential effects of shopping centers and several other activities on daytime population distributions. The INE (2003a) study, by listing the number of people commuting by origin/destination (O/D), indicates the total number of residents in each municipality who leave their homes daily. The difference from this figure to the total resident (i.e., nighttime) population, constitutes the daytime residential population (see Table 8). This total is then transformed in a ratio that expresses the proportion of the resident nighttime population that does not leave home during the day. By multiplying the nighttime population grid by this ratio, the daytime residential population grid is obtained.

Table 8. Resident population and figures calculated from O/D matrix in the study area, 2001.

<b>Municipality</b>	<b>Resident population</b>	<b>Daytime population</b>	<b>Population originating in the municipality</b>	<b>Population destined to the municipality</b>	<b>Daytime residential population</b>
Cascais	170,683	151,115	87,056	67,488	83,627
Oeiras	162,128	148,937	84,777	71,586	77,351

*Source: INE, 2003a*

Adding the total daytime residential population in each municipality to the people whose destination is that municipality (whether originating outside or inside) yields the daytime population, which in the study area is lower than the resident population (see Table 8). This population whose destination is the municipality constitutes the total workforce and students present during the daytime - the workforce population.

This figure was used as a control to “calibrate” the detailed data on workforce and students, since the INE study was assumed to provide the official and most reliable reference population totals for this project. In order for the total workforce population obtained from disparate sources to agree with the figures from INE, it was necessary to scale the DGEEP data. This was accomplished by approximation: by querying the databases it was discovered that the sum of school population added to the sum of personnel of workplaces with five or more employees would approximate the INE value. For a perfect match final adjustments were made by adjusting the value of workforce of those few public services for which specific figures were unavailable.

With this option it was assumed that the personnel of private workplaces with up to four employees worked in, or in the close vicinity to, the place of residence (e.g., ground floor of multi-story building). This solution is simplistic but correlates well with the empirical notion that this situation

frequently occurs in smaller businesses in the study area (e.g., small retail shops, bars and cafés, etc.). Regarding public workplaces it was assumed that these do not correspond to anyone's residence and all employees have to commute to work.

The (daytime) workforce population grid was created by georeferencing all workplaces (private and public) and schools in the study area. This phase was the most labor-intensive and time-consuming of the whole project, owing to three main reasons: a) the poor quality of address information in the DGEEP database and model requirements, b) the need to complete the database by obtaining data for public services from many disparate sources, and c) the presence of firms and activities that due to temporary nature had no address specified below the municipality identifier.

Due to the first factor it was necessary to manually geo-reference 1,395 workplaces (32% of the total), because details of address and/or reference information were insufficient to allow geocoding in ArcGIS with the positional accuracy required by the high resolution model. Such was the case of shopping centers, whose large size required that individual stores were manually located in their interior. This was done in ArcView using as main reference the street centerlines and 1-meter orthophotos for 1995, with the aid of available maps and up to date high-resolution imagery from Google Earth. Fieldwork was also carried out to locate some workplaces, especially in rural areas and industrial areas, and to confirm the location of others.

The second problem was overcome by identifying all public services in the study area and through personal contacts by email or phone requesting information on the number of staff for 2001. Using this process, an additional 212 public services and staff were identified in the study area. Data on the 129 public schools existing in 2001 were obtained in a shapefile from the Lisbon Metro Area, but it only contained information on the number of students, teachers, and staff for Cascais. The number of students per school in Oeiras was obtained from the Ministry of Education website and values for Cascais were then used to calculate teacher and staff-to-student ratios to proportionately estimate counts of teachers and staff for Oeiras.

The third problem, accounting for temporary workers without a permanent work address, could not be resolved in a systematic way. Therefore, it was decided to randomly distribute them to located workplaces in the respective municipality. This was accomplished by using a random number generator in Excel to select as many workplaces as the total number of temporary workers and then increasing their staff by one. In this fashion, 1019 temporary workers were distributed in Oeiras, and 812 in Cascais. Although not entirely satisfactory, this simplification is likely to have low impact in overall results since it only affects 1.4% of all workers present in the study area.

All workplaces for which complete and reliable addresses were determined, 2921, were geocoded in ArcGIS. Geocoding is a semi-automatic

process that assigns coordinates to an address by comparing these descriptive elements to those present in the reference material. Street centerlines obtained from GeoPoint provided the reference used for geocoding workplaces' addresses. Geocoding was carried out at several levels and in an iterative way. First, addresses having ZIP code 4+3 (*código postal*, in Portugal) were geocoded to one street side; then, remaining addresses were geocoded to the other side of the street; finally, remaining addresses having only ZIP code 4 were geocoded. Figure 13 shows all workplaces manually georeferenced and geocoded in the study area. Finally, all georeferenced workplaces were converted to a 25-m grid using the number of workers or students for each cell value, resulting in the (daytime) workforce population grid.

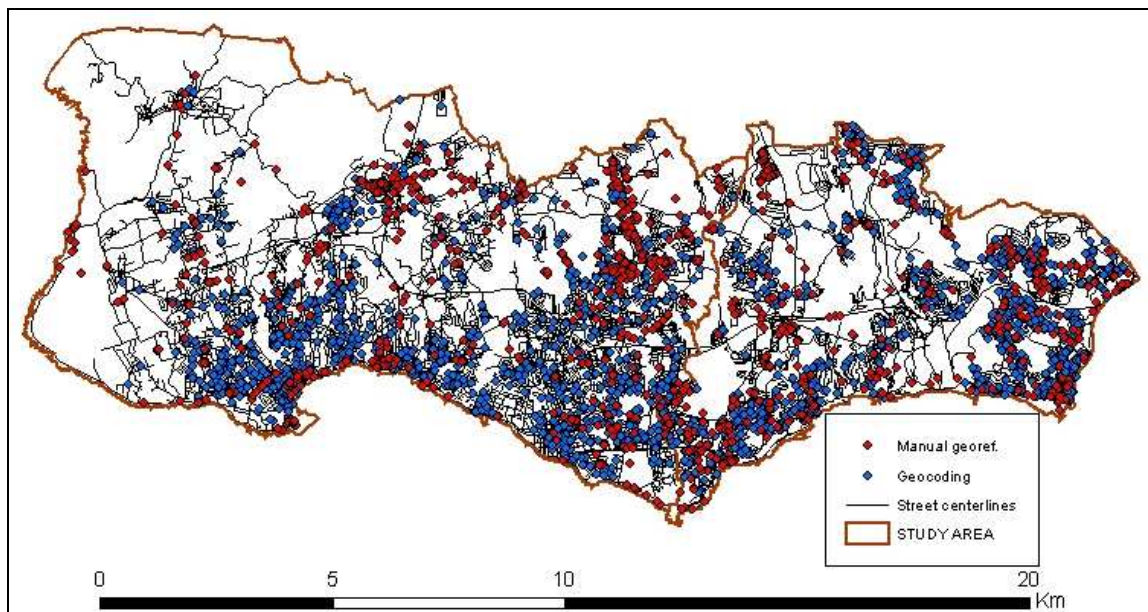


Figure 13: Places of work and study georeferenced in Cascais and Oeiras.

### 3.7 Ambient Population Distribution

The ambient population aims to represent, using a single measure, the weekly average distribution of population considering where the main human activities of sleep, work, and study take place.

The ambient population distribution is estimated by computing a weighted average of nighttime and daytime distributions, considering the proportion of nighttime and daytime periods occurring in a 7-day weekly cycle. The grid of ambient population was computed according to equation 3:

$$AP_i = (NRP_i * 9/14) + (DP_i * 5/14) \quad (3)$$

where  $AP_i$  is ambient population for each cell  $i$ ,  $NRP_i$  is nighttime residential population for each cell  $i$ , and  $DP_i$  is daytime population in each cell  $i$ . Because the working schedules tend to be long and highly irregular in metro areas in Portugal (morning rush hour usually starts at 7:30 AM and evening ends as late as 9:00 PM) a simple division of week days in two twelve-hour periods of work and rest was assumed. Of those total fourteen day and night periods occurring in a typical week, daytime population distribution will occur in five of those (i.e., daytime workdays), while the remaining nine periods will be better represented by the nighttime distribution.

### 3.8 Verification and Validation

For a rigorous quantitative assessment of model results, data are required which represent 'ground truth' with adequate accuracy and compatibility with the modeled data. However, a frequent driver for projects such as this one is precisely the absence of a suitable reference database, a fact that limits the scope of validation (Dobson, 2000; McPherson and Brown, 2003).

Therefore, validation of daytime population distribution was limited by unavailability of compatible reference data sets for Portugal. However, it can be argued that this limitation is less relevant since daytime population distribution originates from a combination of the nighttime distribution surface (subject to formal validation) with mostly official statistical data, as opposed to being derived from heuristic or empirical weights. Still, the daytime population distribution was subject to verification in several ways:

1. The input data (especially workplaces' addresses from DGEEP) were verified through cross-checking with other sources and field work, regarding location of workplaces (firms).
2. Results were verified with high-resolution imagery to confirm positional accuracy of distributions.



3. A check that the total number of workers provided and other statistics (census and mobility) were not contradictory. The specific number of workers and students obtained by workplace and school, the official census counts, and the mobility figures were not questioned or verified.

Nighttime population distribution was subject to a formal accuracy assessment process, using the higher-resolution of census blocks (i.e., *subsecção*) as reference (i.e., ground truth). However, there is no standard or consensus method for quantitative assessment of interpolated population distributions, and different accuracy measures have been used by different authors: mean percent error (MPE) and mean absolute percent error (MAPE) (Goodchild et al., 1993), root mean squared (RMSE) error (Fisher and Langford, 1995), RMSE and coefficient of variation (CV) of RMSE (Eicher and Brewer, 2001), and correlation analysis (McPherson and Brown, 2003). Therefore all of these measures were computed for each municipality in the study area.

Cell values in modeled distributions were aggregated by census block in ArcGIS and compared against the respective census count, and accuracy measures were calculated in Microsoft Excel. For each municipality, scatter plots of correlation analysis were also created, and the MPE mapped.

## 4. RESULTS AND DISCUSSION

In this section, model results are presented and discussed. Main model results comprise floating-point raster grids of nighttime population distribution, daytime worker and student population distribution, daytime residential population distribution, and total daytime population distribution. These results represent, for the study area, maximum expected population densities in 2001 for each 625 square-meter grid cell.

Grids approximating ambient population distribution considering a weekly cycle are also computed and presented.

### 4.1 Nighttime Population Distribution

The nighttime (residential) population distribution grids (Figure 14 and Figure 15) represent maximum expected population density by cell for a typical nighttime period, assuming that all people are in their respective homes. This is obviously a simplification of reality, since on any given night an unknown number of people will not be in their residences for varied reasons, such as nighttime work shifts, travel, leisure, etc. Therefore, resident population density is over-estimated and results misrepresent those activities taking place in the nighttime period and do not account for people that may

occupy hotels and other lodging, either originating from inside or outside the study area.

Use of this raster dasymetric mapping approach allows the allocation of population within source zones (i.e., block groups) to those residential areas where it is really present, at high spatial resolution. However, the use of simple areal weighting for distributing population in each dasymetric zone is subject to the assumption of even density which may not correlate with reality in some areas, such as those areas comprised of both single-family dwellings and multi-storey apartment buildings. Fortunately this situation is avoided in the design of census zones, and especially uncommon at the block level.

Residential streets seem to provide adequate support for disaggregating official census counts, but the nighttime population distributions closely reflect the mapping of residential streets, which in turn is dependent on land use classification as residential. Therefore the approach used here is subject to the combined effects of error from different ancillary data sets and its propagation to the final results. Still, it is believed that results represent improvement over the classical census representation of population, and the raster format offers flexibility by facilitating computation of total counts to any zoning.

Results preserve the overall relative differences in population density between Cascais and Oeiras (Cascais is approximately half that of Oeiras)

and reveal a pattern of significant range of nighttime densities within the study area, especially in Oeiras.

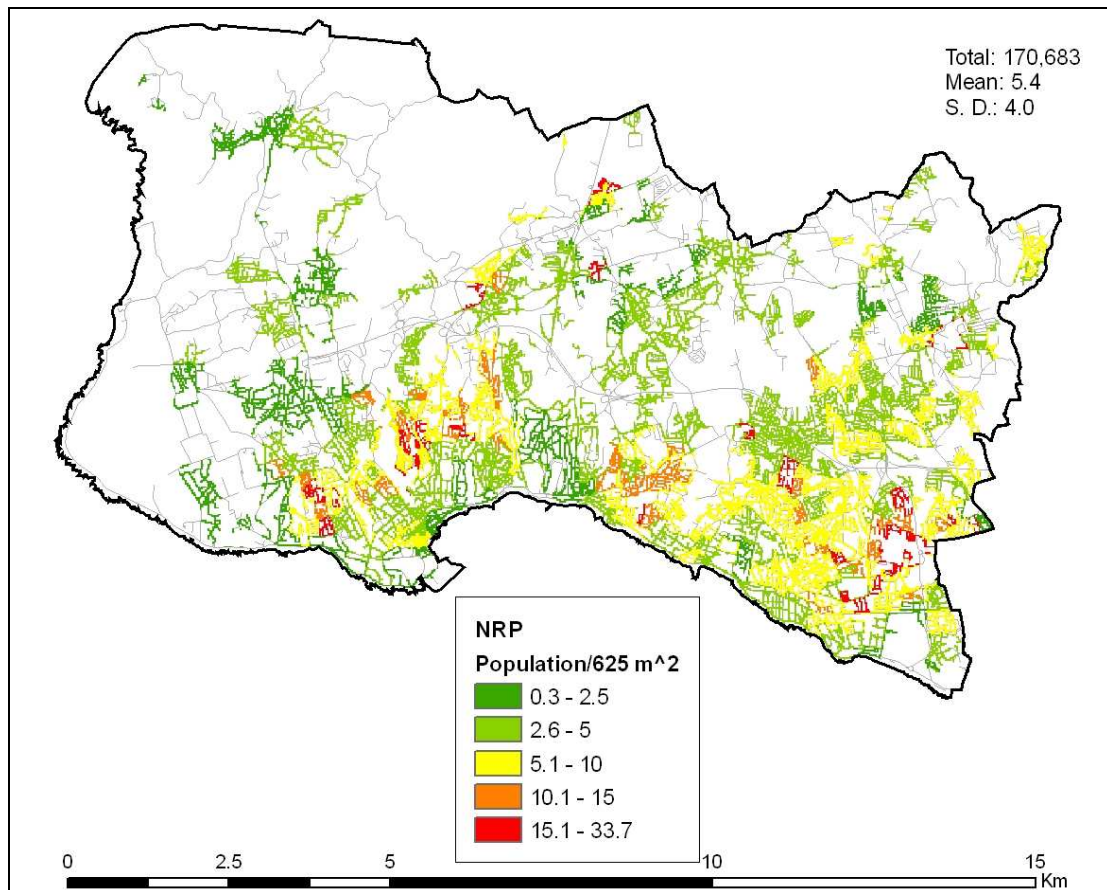


Figure 14: Grid of nighttime population distribution in Cascais.

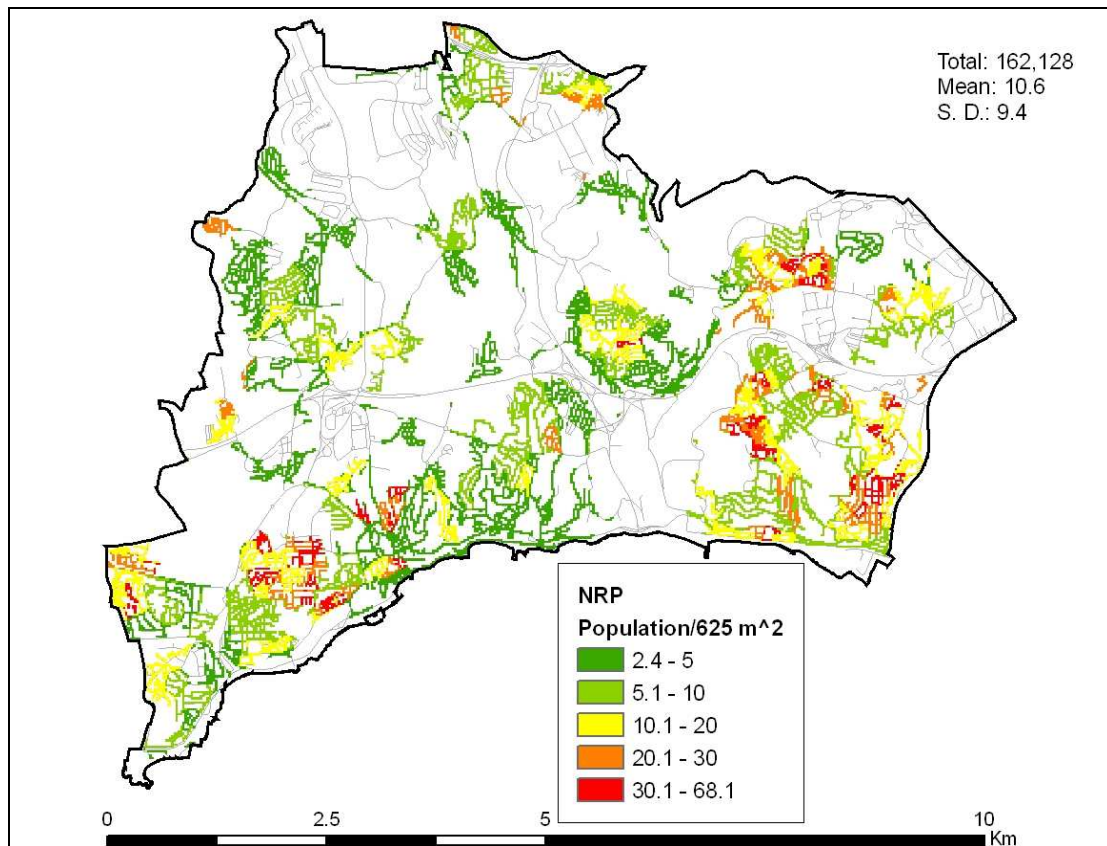


Figure 15: Grid of nighttime population distribution in Oeiras.

#### 4.2 Daytime Residential Population Distribution

The daytime residential population distribution grids (Figure 16 and Figure 17) is the residential component of the overall population distribution during daytime hours, and represents the maximum expected residential population density by cell during a typical daytime period. This study assumes that all people who do not leave for work or school remain in their homes. In reality, an unknown number of people will be involved in regular activities taking place away from home for at least part of the day (e.g., shopping,

leisure, sport practice etc.). This assumption causes daytime residential population densities to suffer from over-estimation.

A second bias in this dataset is introduced because estimation of residential population for each municipality is dependent on a single ratio that expresses the proportion of the nighttime (resident) population that does not leave home for work or study during the day. Because mobility figures used to derive that ratio are only available at the level of municipalities, it is assumed that this proportion is constant across that area, which may not be true. Ideally, this ratio should be more disaggregated or differences estimated using ancillary variables such as mean age of residents, since older or retired individuals are expectedly more likely to remain home. McPherson and Brown (2003) also report the same limitation in their model.

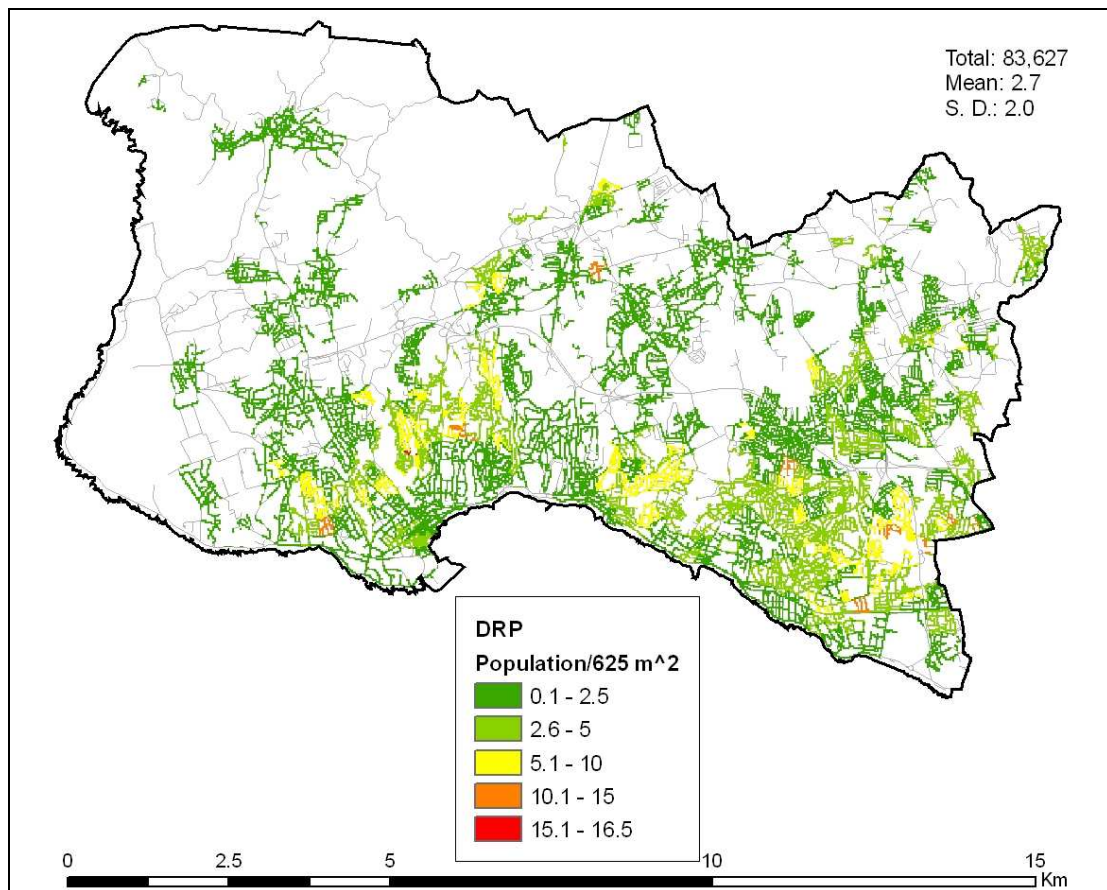


Figure 16: Grid of daytime residential population distribution in Cascais.

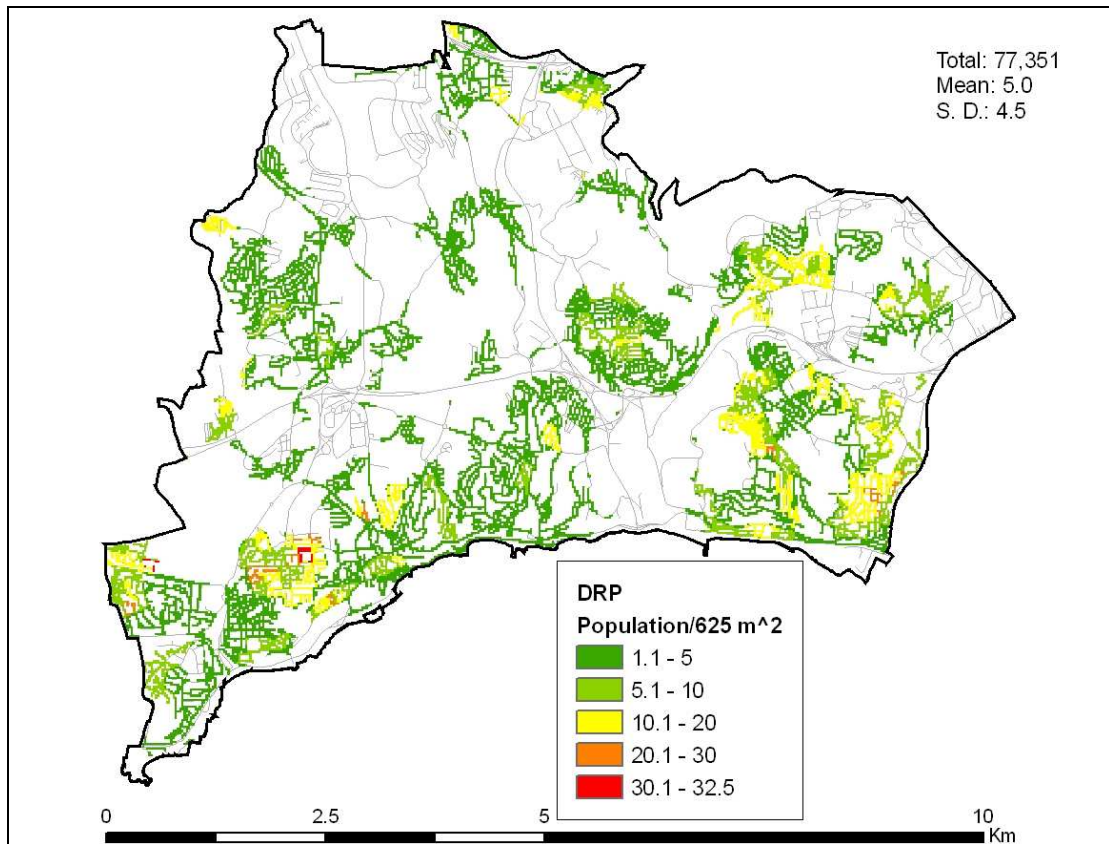


Figure 17: Grid of daytime residential population distribution in Oeiras.

#### 4.3 Daytime Worker and Student Population Distribution

The daytime worker and student population distribution grids (Figure 18 and Figure 19) is the displaced component of the overall population distribution during a typical daytime period, and represents maximum expected worker and student population density by cell during the daytime period, assuming that all workers and students are present in their listed workplaces or schools.



In reality not all students will be in school all day, and on any given day an unknown number of workers will not be in their listed workplaces. Absent workers may be on sick leave, traveling, visiting customers, etc. But even if they are at work, for some workers their actual workplace will be different from their listed workplace. This situation occurs in 'distributed activities' (such as security, cleaning, transportation, etc.) where workers are either traveling (i.e., transportation) or are commonly assigned to workplaces of customers of the firm for which they are listed as staff (i.e., security and cleaning). Although agriculture is a residual activity in the study area, the model also misrepresents presence of farm workers in agricultural fields. Some densities are also over-estimated because the grid cell size is too small to represent the actual area occupied by some large workplaces and schools. Use of parcel data or some other information relating to building footprint in geocoding and converting the whole building footprint and associated population to several grid cells would better model these situations. These factors contribute to the occurrence of a wide range of densities in the study area, especially in Oeiras.

For dissemination, aggregating the original results to a lower-resolution would generalize results by reducing the variability of cell values towards their mean, while still preserving usefulness. For example, 50-m grids would still be fine enough to support analysis at the local level, but would reduce the

problem of a work place being represented by a grid cell smaller than the footprint of the building housing the workers.

Especially important for the quality of these distributions is the positional accuracy of workplace locations, and therefore significant effort was devoted to verify and ensure their accurate geo-referencing.

The mean density value of these distributions does not represent average business size since several workplaces may occupy the same grid cell.

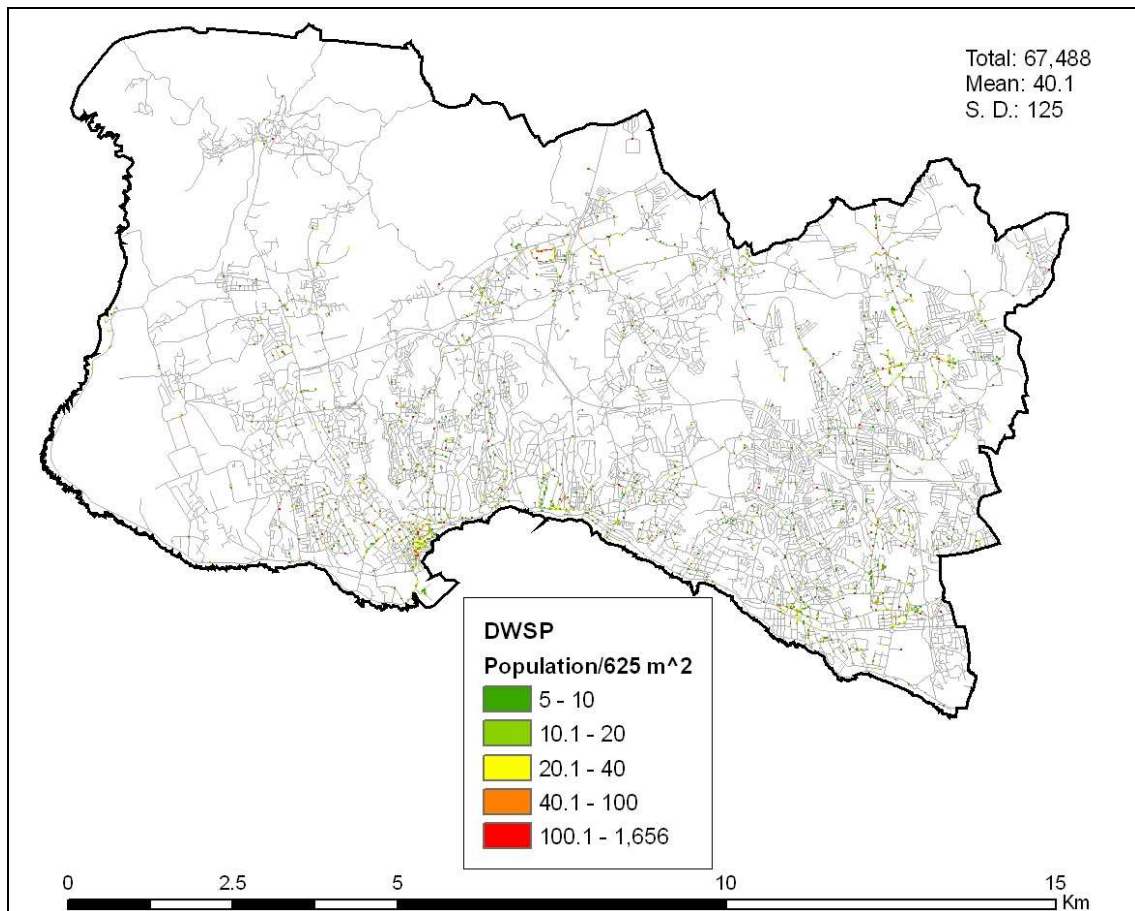


Figure 18: Grid of daytime worker and student population distribution in Cascais.

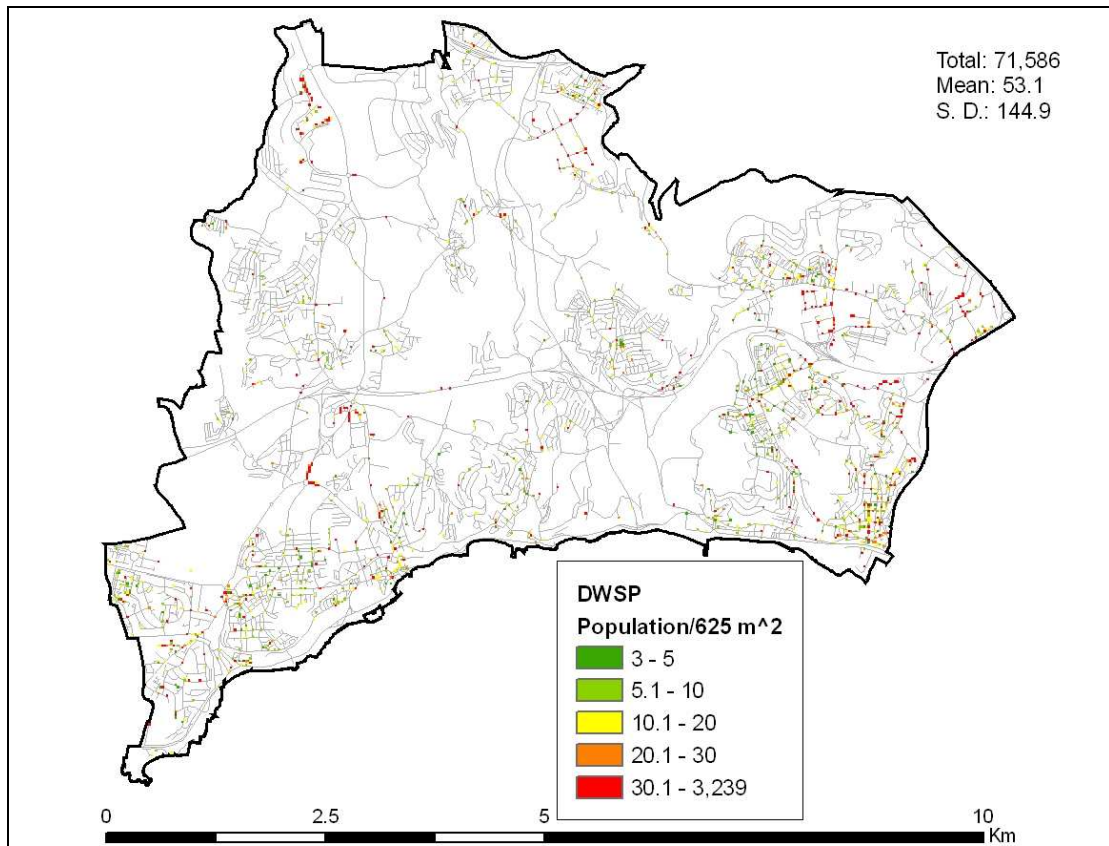


Figure 19: Grid of daytime worker and student population distribution in Oeiras.

#### 4.4 Daytime Population Distribution

The daytime population distribution grids (Figure 20 and Figure 21) represent maximum expected overall population density by cell in daytime period assuming that all people are present in their assigned workplaces or school locations and the rest remains at home.

Resulting from the simple arithmetic sum of the daytime residential population distribution with the daytime worker and student population

distribution on a cell-by-cell basis, this representation inherits the limitations and caveats of the input datasets which were outlined above. Still, the daytime distribution has the merit of adequately accommodating people working at home, when compared with an approach employing figures of active population and assuming that those workers regularly commute.

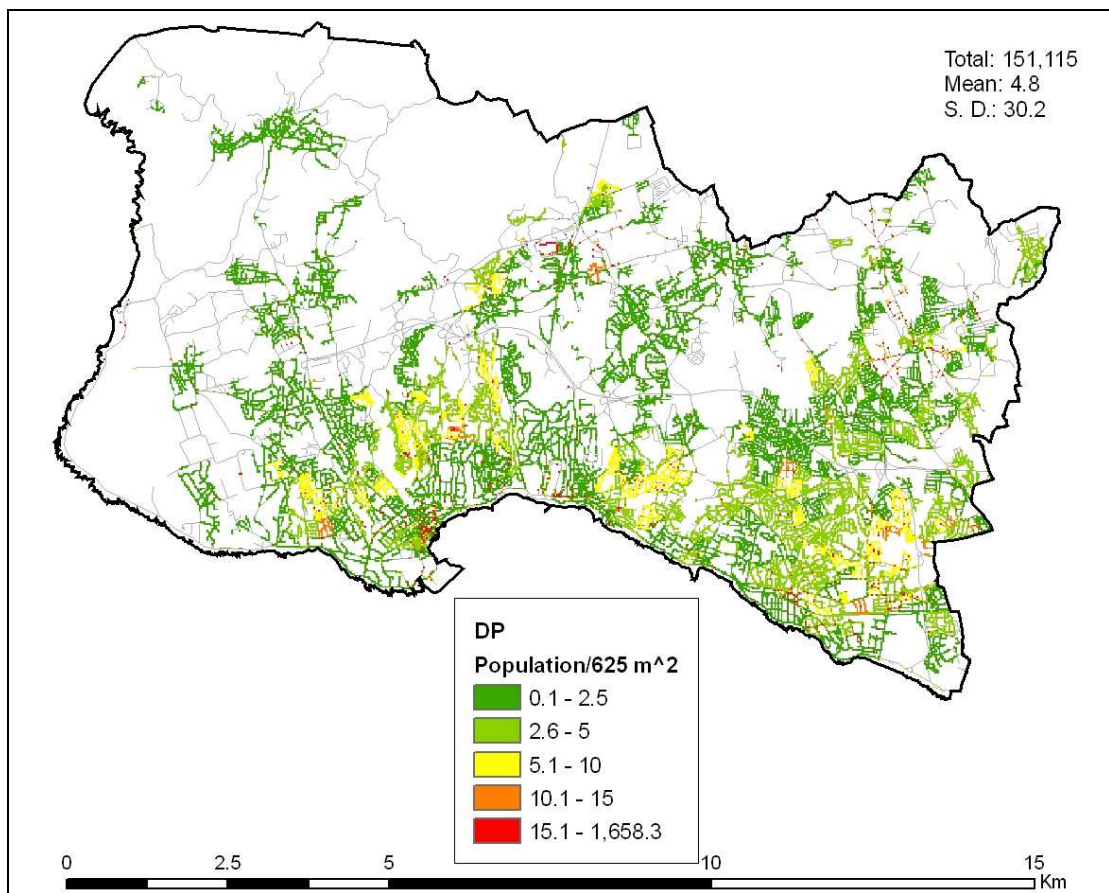


Figure 20: Grid of daytime population distribution in Cascais.

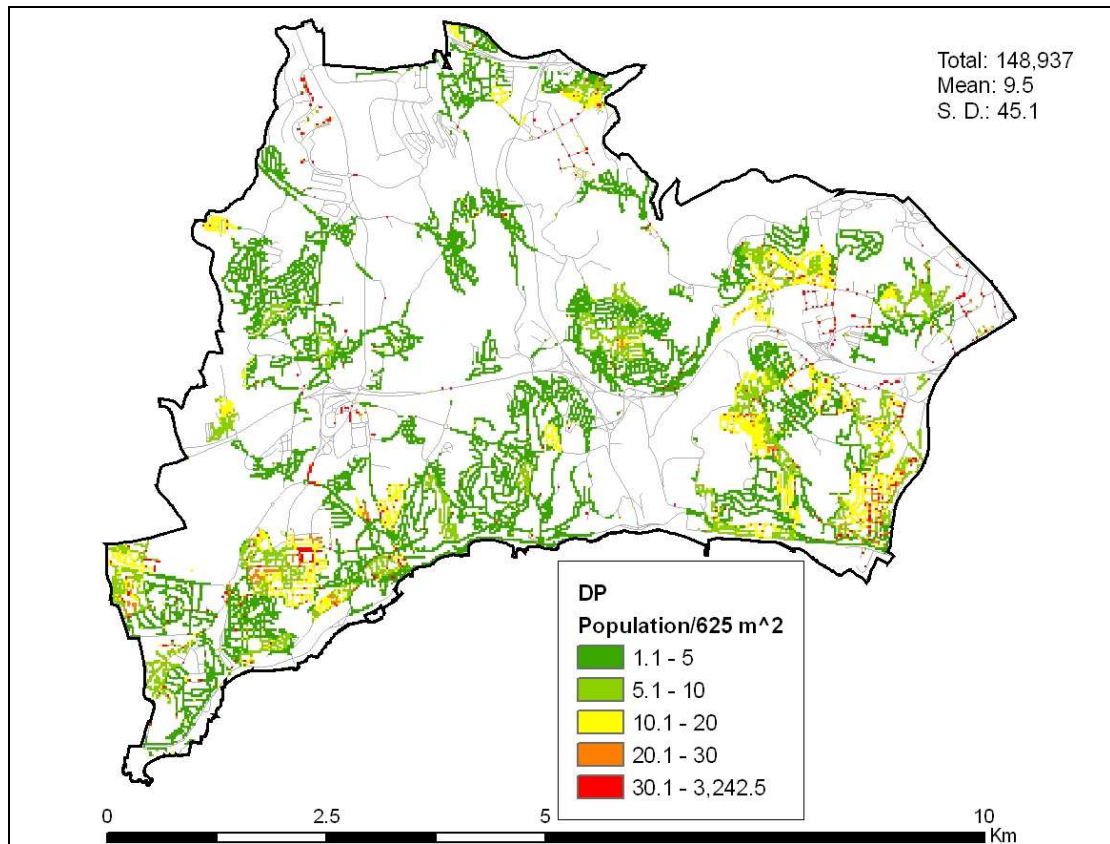


Figure 21: Grid of daytime population distribution in Oeiras.

#### 4.5 Ambient Population Distribution

The ambient population distribution grids (Figure 22 and Figure 23) represent an approximation to a weekly average distribution of population density considering where the main human activities of sleep, work, and study take place.

Resulting from the arithmetic weighted sum of the daytime residential population distribution with the daytime worker and student population distribution on a cell-by-cell basis, this representation inherits the limitations

and caveats of those input datasets. Namely this representation does not account for people involved in activities different than the main ones cited above, but it would be important to represent people present in transportation networks and shopping centers because there is the perception that on average these are frequented by a significant number of individuals in the study area. It would be especially important to represent these activities in finer temporal segmentations that would represent periods of intense commuting (i.e., 'rush hour') or leisure (e.g., weekends).

Although available census and commuting statistics for the study area do not account for population shifts other than for work or school, an ambient population map would greatly benefit from inclusion of important fluxes such as those caused by tourism, especially in areas where tourism influx significantly increases counts of actual population present. Inclusion of daily shopping activities based on known distributions of shopping centers and seasonal tourism activities could be included on a probabilistic simulation, but would require additional information to be used in a similar way to the commuting statistics used in this study. However, this project also demonstrates that modeling the temporal distribution of population becomes increasingly difficult and challenging as spatial resolution increases (Sutton et al., 2003), while limitations of source data and results become more apparent.

Perhaps the representation of ambient population could be further improved by smoothing values using a mean spatial filter that simulates



people's movement in the vicinity of home, workplace or school, as suggested by Sutton et al. (2003). To produce realistic results this approach requires knowledge of average distance traveled and information that this distance be rather invariable throughout the study area. However, this information would have the advantage of preserving the model's spatial resolution.

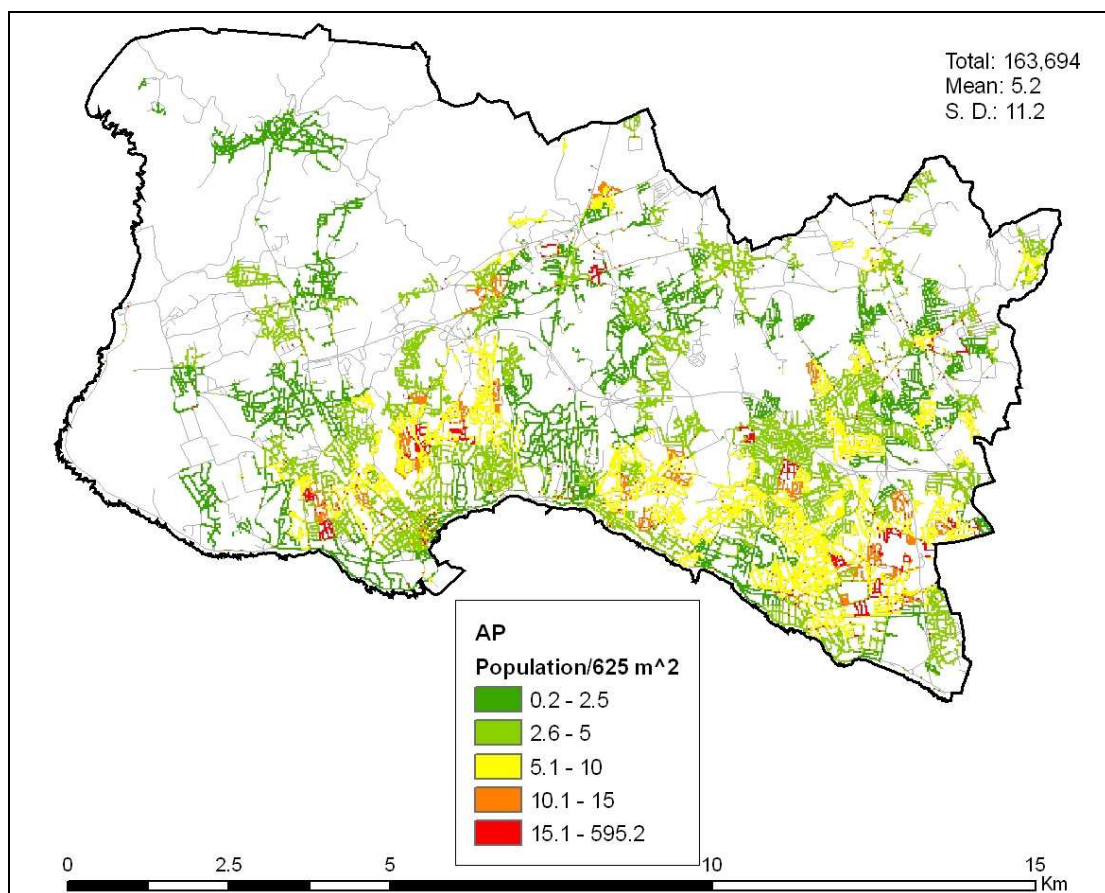


Figure 22: Ambient population distribution in Cascais.

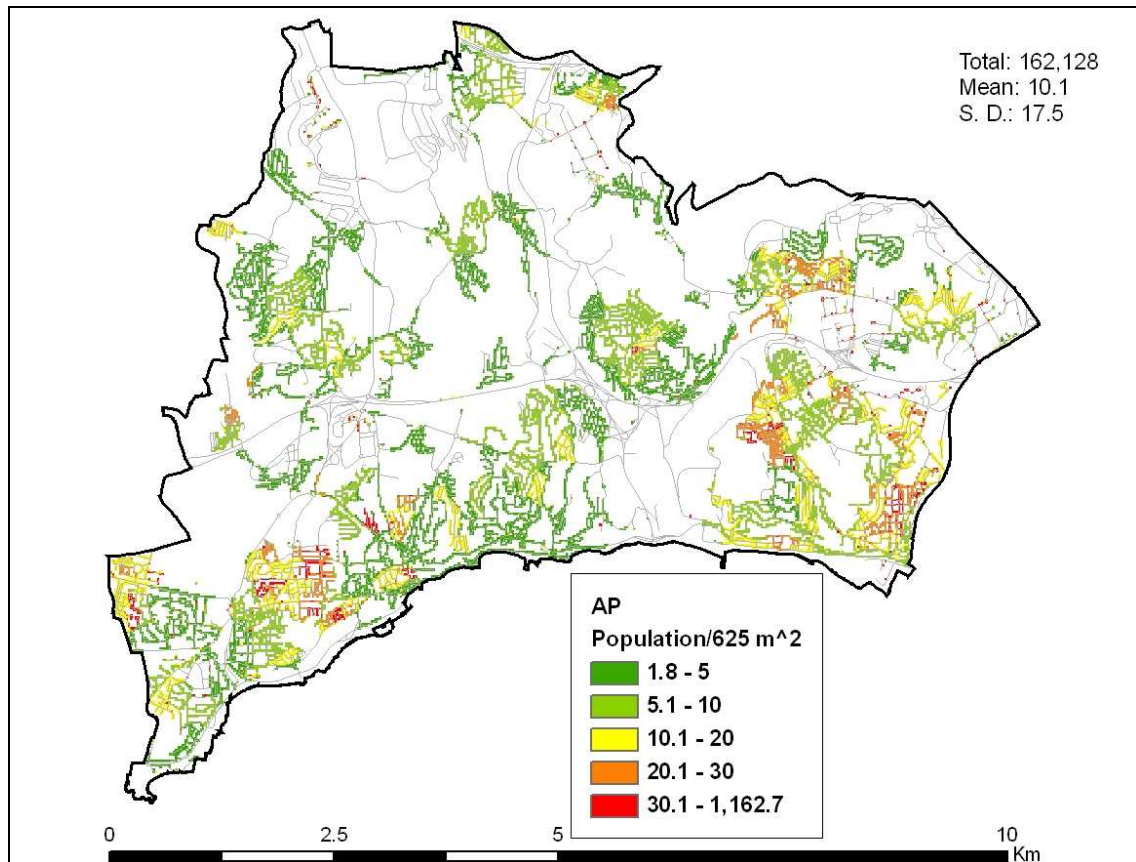


Figure 23: Ambient population distribution in Oeiras.

#### 4.6 Verification and Validation

There is no standard or consensus method for quantitative validation of modeled population surfaces, but the nighttime population distributions were assessed using different accuracy measures which have previously been employed in similar studies.

Because zone-based modeling was adopted at the level of census block groups (source zones), and despite results being presented in raster



format, it seems appropriate that the dasymetric distributions be assessed using the smaller census blocks. In contrast, results of a cell-based model would be best assessed against a reference dataset that had the actual population by cell at the same resolution as the model. However, such reference data sets are usually non-existent and accuracy values are very dependent on positional accuracy with which model and reference data set are co-registered (Sutton et al., 2003), leading to preference for zone-based assessment even in these cases.

In the present case, since results are modeled at the level of the census block group and assessed at the lower level of the census block, it is important to characterize these enumeration units in the study area (Table 9).

Table 9. Characteristics of census blocks and block groups in Cascais and Oeiras.

Municipality	Blocks			Block groups		
	No.	Area (ha)		No.	Area (ha)	
		Mean	S.D.		Mean	S.D.
Cascais	2433	4.0	2.2	289	33.6	91.3
Oeiras	1486	3.0	6.4	240	19.1	37.4

Table 9 shows that there is a large number of census blocks in each municipality, which allows for parametric analysis of results using blocks as observations. It also shows that census blocks are on average quite small and more homogenous in size in Cascais than in Oeiras, while block groups are very heterogeneous, more so in Cascais.

Table 10 shows the mean and standard deviation (S. D.) values of block-level population in census and model data for the study area.

Table 10. Population statistics by block for Cascais and Oeiras.

<b>Municipality</b>	<b>Census</b>		<b>Model</b>	
	<b>Mean</b>	<b>S.D.</b>	<b>Mean</b>	<b>S.D.</b>
Cascais	70.2	98.3	70.2	89.5
Oeiras	109.1	128.8	109.1	116.9

Using the values obtained by assessment at the level of the census block, overall error measures were computed for each municipality in the study area and are presented in Table 11.

Table 11. Overall accuracy measures for Cascais and Oeiras.

<b>Municipality</b>	<b>RMSE</b>	<b>CV</b>	<b>Measure</b>		
			<b>r</b>	<b>MPE</b>	<b>MAPE</b>
Cascais	53.7	0.77	0.84	77.9	114.0
Oeiras	80.7	0.74	0.79	147.2	183.3

Values show that the model generally performed better in Cascais according to every error measure, with the exception of coefficient of variation (CV). CV is a standardized value obtained by dividing the root mean squared error (RMSE) by the actual mean value for the municipality. RMSE is a useful measure because it can be applied to count data and has the same units as the mapped variable (Eicher and Brewer, 2001). Although the RMSE is higher for Oeiras, the CV is lower due to the fact that the actual mean population per

block is higher than in Cascais (109.1 vs. 70.2). This means that although nominally a larger number of people are on average being misplaced in Oeiras, this is less significant in distorting the overall distribution than it is in Cascais.

Following the approach taken by McPherson and Brown (2003), correlation analysis was also conducted to assess model performance. Bivariate correlation analysis evaluates the strength of association between two variables (Burt and Barber, 1996), in this case between the same two variables drawn from different data sets -- modeled counts and actual census counts in each block. The Pearson's product-moment correlation coefficient ( $r$ ) was also computed, the most commonly used measure of correlation for interval or ratio variables (Burt and Barber, 1996). Scatter plots of the correlation analyses for Cascais and Oeiras are presented in Figure 24 and Figure 25, respectively. The charts show that the model has a tendency to underestimate in census blocks with intermediate population totals, i.e., those having from 100 to 400 people. This effect is most probably due to the fact that residential street density does not adequately represent population density in those areas especially because the study assumes constant density in dasymetric zones.

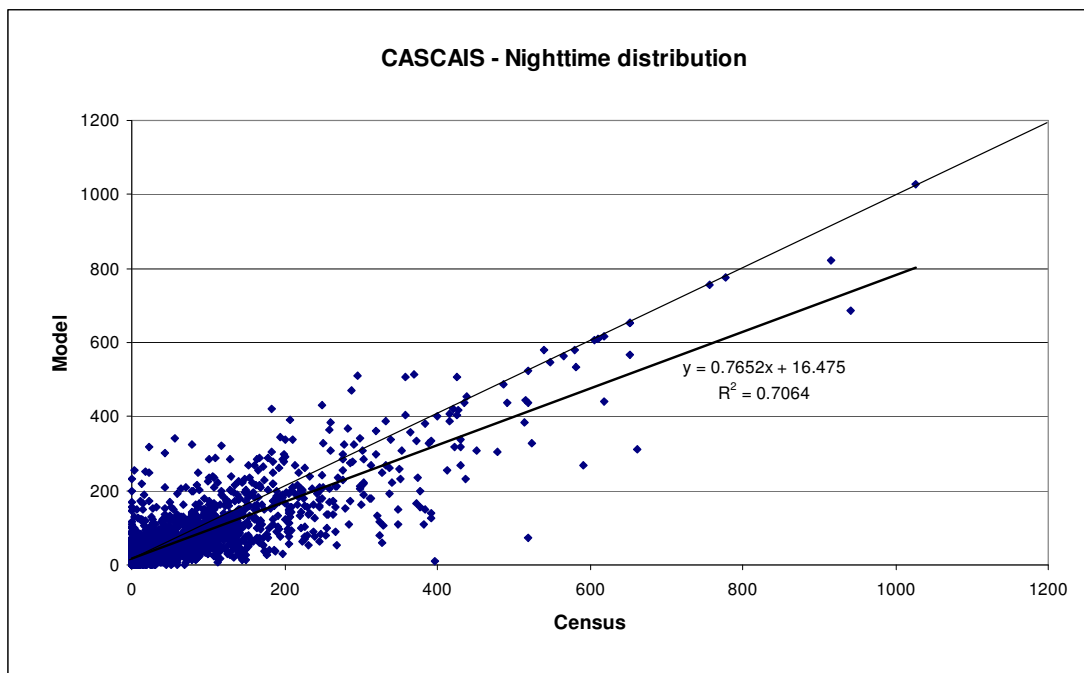


Figure 24: Comparison of census vs. modeled population by block in Cascais.

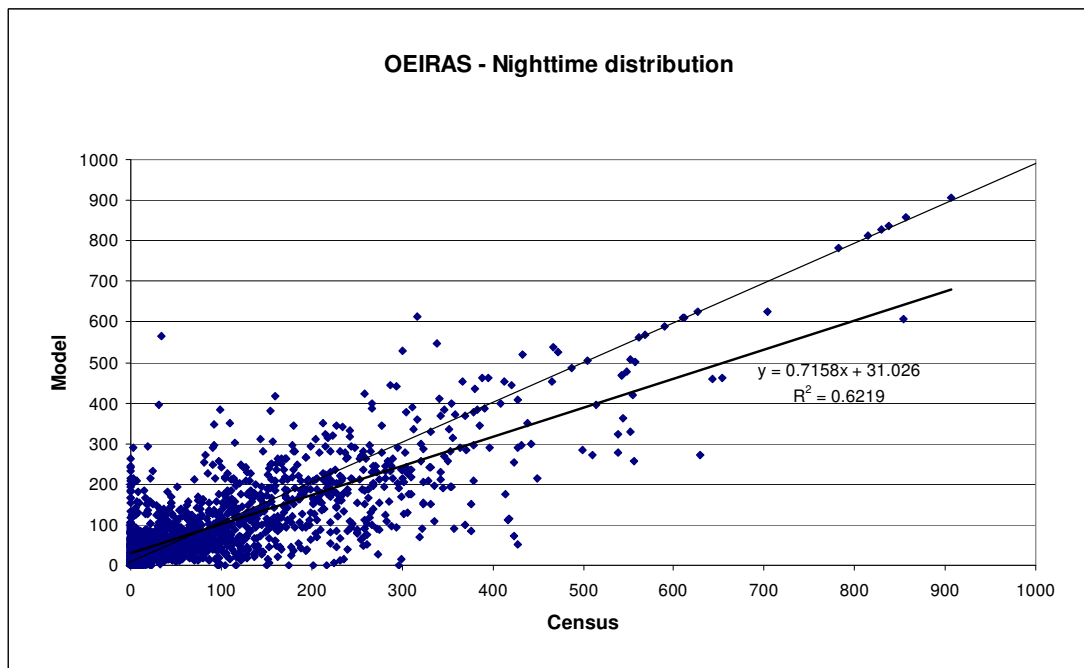


Figure 25: Comparison of census vs. modeled population by block in Oeiras.

This analysis can be summarized by the correlation coefficient, a dimensionless variable that equals 1 when the association is perfect, i.e., there is an increasing linear relationship. The obtained overall correlation coefficients for each municipality were both high, but with a higher value for Cascais (0.84) than for Oeiras (0.79), indicating a good fit between model results and reality. Even when the 36 blocks that coincide with block groups are not considered, the coefficients decrease only slightly to 0.79 and 0.75, respectively.

In accordance with Goodchild et al. (1993), mean percent error (MPE) values for each municipality were obtained by computing the error for each block as a percentage and calculating its average, while the mean absolute error (MAPE) was obtained by averaging the absolute percentile error by block. An overall MPE for each municipality has little explanatory power for errors, mostly revealing if there is a tendency for under or over-prediction of population by the model and its intensity. Obtained MPE values show that over-prediction dominates errors by blocks, and that it is stronger in Oeiras. Municipal MAPE values show that on average, by census block model predictions are off by 114% in Cascais and by 183% in Oeiras. However, MPE and MAPE can be deceiving measures to assess overall model quality because all samples (blocks) carry the same weight in the calculation

regardless of the number of people actually mis-estimated or the population in the block (error count).

Percentage error may be a measure more useful for mapping and analyzing the spatial distribution of error, and therefore the percentage errors calculated for each of the 3,919 census blocks in the study area were grouped in five classes and mapped (Figure 26).

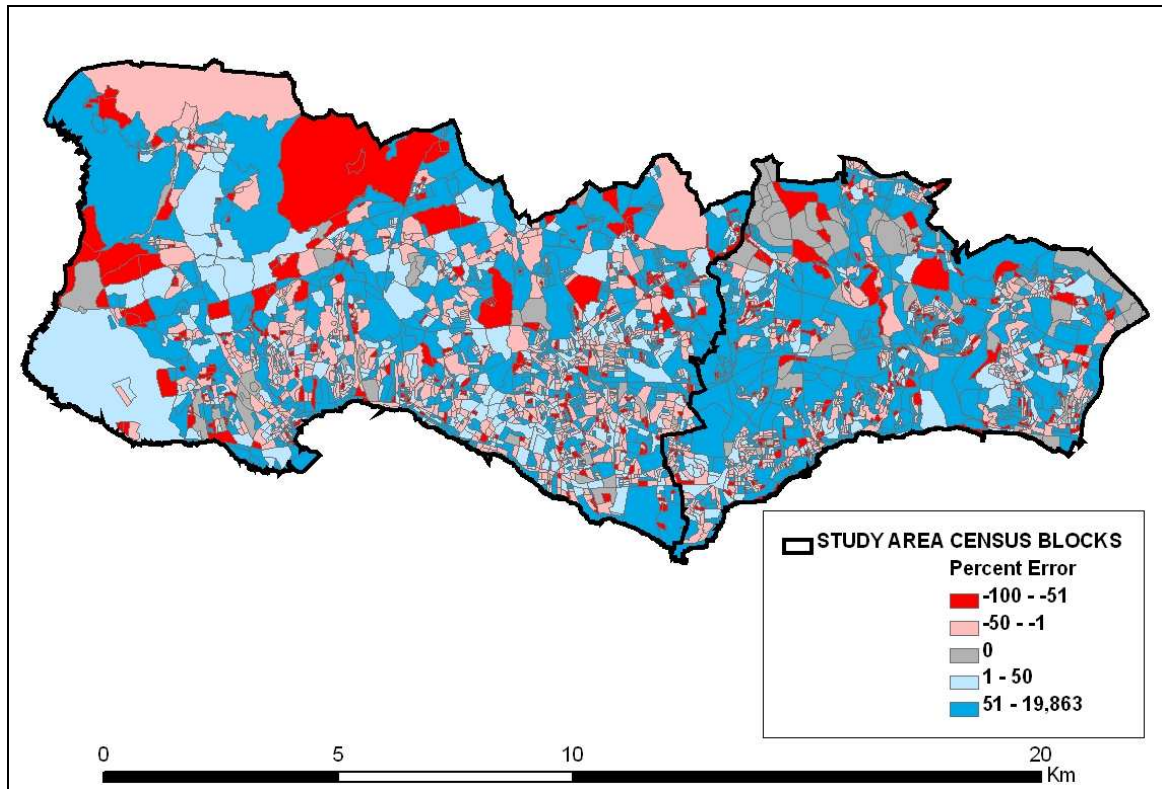


Figure 26: Map of percentage error by census block in Cascais and Oeiras.

Although Figure 26 does not reveal a clear spatial pattern of error, a brief statistical analysis shows that while population is over-predicted for the majority of blocks, as previously stated, blocks where the population is under-predicted account for the majority (63%) of the population in the study area (see Table 12). The population is significantly over-predicted in a large number of blocks (1,318) that represent only 11% of the total population, while population is significantly under-predicted in 573 blocks that represent 19% of the population. In short, the model tends to under-predict population in smaller, more urban blocks with high densities and over-predict in larger rural blocks with low densities, but significant mis-estimation affects only 30% of the total population in the area.

Table 12. Statistical characterization of percent error by census block in the study area.

<b>PE Class</b>	<b>No. of blocks</b>	<b>Mean Area (ha)</b>	<b>Mean</b>	<b>Population Total</b>	<b>%</b>	<b>Mean Pop. Dens (sq km)</b>
-100 to -51	573	3.2	110	63,014	19.0	14,774
-50 to -1	1,152	2.4	127	145,946	43.9	11,726
0	141	6.9	161	22,276	6.7	7,362
1 to 50	735	3.7	89	65,237	19.6	8,152
51 to 19,863	1,318	4.6	27	35,888	10.8	2,419

To allow for cartographic depiction of the quality of modeling by source zone (census block group) the absolute percentage errors calculated for each

census blocks were averaged by block group, then grouped in four classes and mapped (Figure 27).

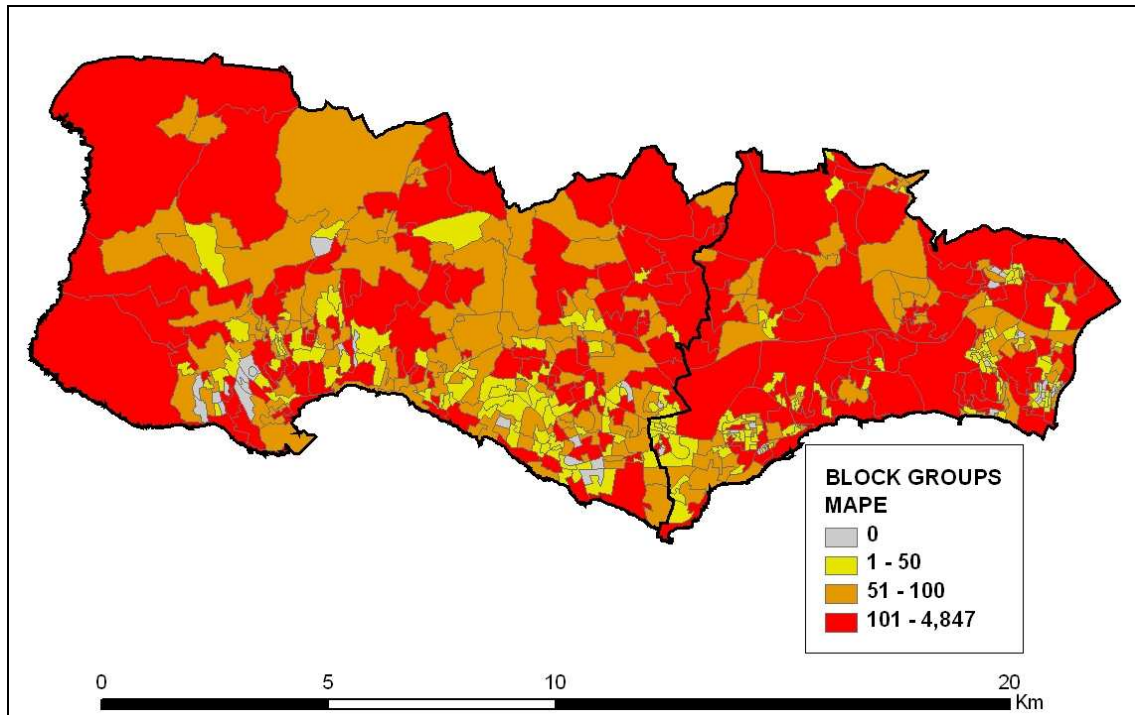


Figure 27: Map of mean absolute percentage error (MAPE) by census block group in Cascais and Oeiras.

Figure 27 shows that on average higher percentile mis-estimations tended to occur in larger, more rural block groups while model results tend to be more accurate in smaller, more urban block groups. However, it should be noted that averaging percentage errors by block group can also be misleading because the number of blocks in each block group varies between



1 (36 blocks) and 36 (1 block) and all blocks are given the same weight regardless of the number of people actually mis-estimated.

Probably a more useful indication for the user of the data is to provide an error measure using the same units and spatial basis as the model results, i.e., population by grid cell. Therefore a count error by modeled cell in each census block was computed and mapped for each municipality, using five classes (Figure 28 and Figure 29).

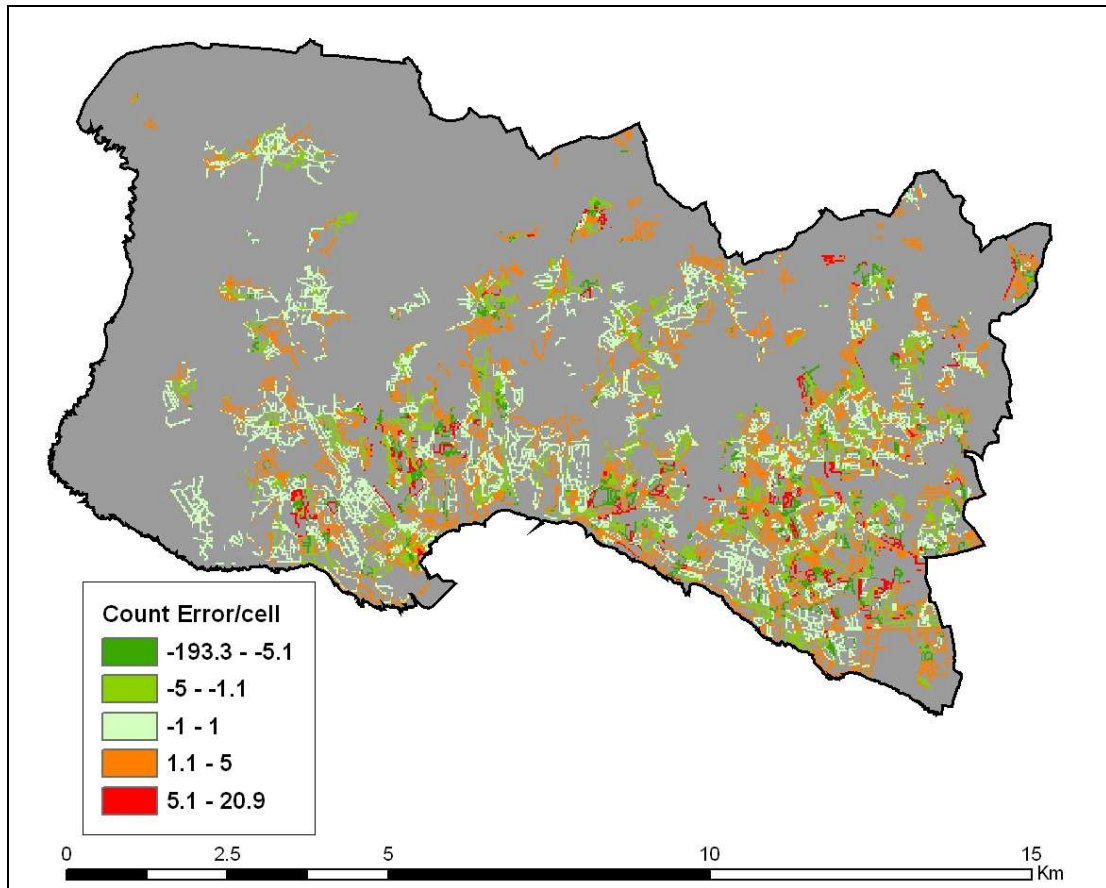


Figure 28: Map of count error by model cell for nighttime distribution in Cascais.

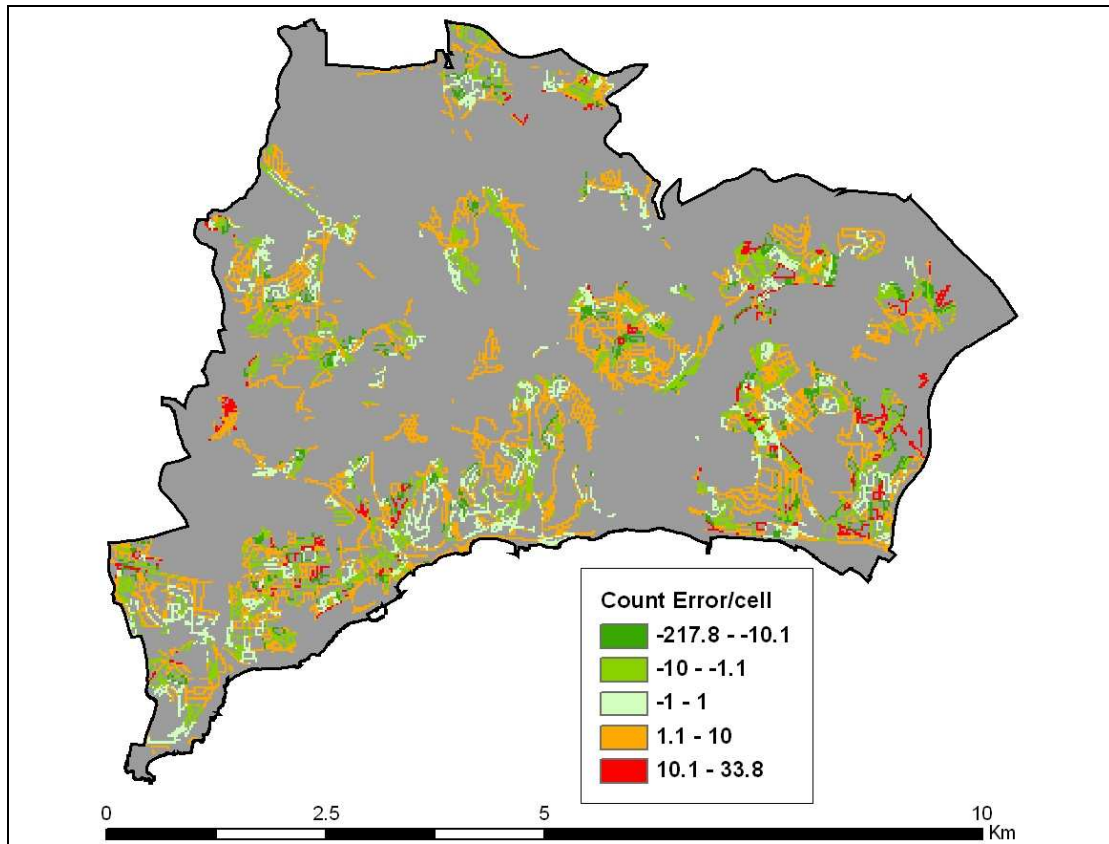


Figure 29: Map of count error by model cell for nighttime distribution in Oeiras.

Cell values represent the difference in number of persons to the value necessary to correctly account for population of the block where they are contained (evenly distributed to all grid cells in a block). Error is not depicted for areas of blocks where census population was present but was not mapped by the model – an error of omission. The model allocated zero population to 84 blocks that were populated (with a total of 2,518 people) according to the census, and inversely allocated population (12,814 people) to 292 blocks that did not have any in the census. Similar to previous results, the range of count errors by modeled cell is higher in Oeiras than in Cascais.

Verification and validation efforts show that overall model performance was very satisfactory given its very high spatial resolution and the rigorous validation conducted using a very high number of samples and varied measures. Eicher and Brewer (2001) obtained 0.73 as the overall mean of coefficient of variation for the binary method, using 1 km<sup>2</sup> cells. Dobson et al. (2000) had to limit validation of their ambient population data to selected areas using residential counts, given the unavailability of ambient counts. McPherson and Brown compared aggregated 250-m model results to county-level data and obtained a correlation coefficient of 0.99. However, their population source zones for modeling nighttime population (census block groups) are nested in the reference zones for validation (counties), and therefore the difference to a perfect correlation is probably due to effect of gridding. Sleeter (2004) obtained correlation coefficients between 0.80 and 0.88 modeling block-group level residential distributions at 30 m.

Still, it should be noted that comparing the quality of different studies and projects in this field based on quantitative error measures is difficult, since all of them use different input data sets, model at different scales and resolutions, and utilize different reference data sets for validation. Even if the current model somewhat mis-estimates nighttime density at the block level, it is still an improvement over an assumption of continuous and exhaustive population presence throughout space displayed by census choropleth maps.

## 5. CONCLUSIONS

A data-driven model was developed and implemented to map the spatial and temporal distribution of population in two municipalities in Portugal for 2001. An approach based on dasymetric mapping was used to combine existing physiographic and statistical data sets to map daytime and nighttime population densities at high spatial resolution. The model used residential streets as ancillary information, which seem to provide an adequate spatial reference base for disaggregating and mapping population distribution at fine scale; using streets as a common reference base allows for direct comparison of daytime and nighttime population distributions even at high spatial resolution.

The model allows temporal segmentation of population distribution into daytime and nighttime populations for a typical workday cycle, increases spatial resolution of nighttime (residential) distributions compared to census data, and models both worker/student and residential components in daytime distributions. Results are produced in a raster format which facilitates aggregation by any zoning for analysis and supports map algebra operations, such as computing the difference between daytime and nighttime population densities for a given area. Model results and characteristics compare favorably with other population mapping efforts, such as LANL's and LandScan USA: LANL's model has a coarser spatial resolution (250 m) and

does not consider population in schools, but its results have a much wider geographical coverage; LandScan USA also has coarser spatial resolution (90 m) but will consider mobility and prison population, and will model various demographic attributes (age, gender, race).

The very high spatial resolution of model results (25 meters) makes them suitable for local-level analysis, and several applications in the field of emergency management have already been demonstrated in recent presentations (Freire, 2007a; 2007b). Increasing spatial resolution beyond 25 m did not significantly improve accuracy of model results.

The modeling method meets basic requirements for distributing population counts, namely the non-negativity constraints and the pycnophylactic condition. This means that nature of the data (ratio) is preserved, and that population is only re-distributed within small source zones (block groups) without their totals being modified.

In addition to the final population products, the model also yields intermediate data sets previously unavailable in Portugal, such as residential streets. The model also approximates a representation of ambient population on a weekly basis through the combination of daytime and nighttime distributions in a single measure, at a spatial resolution 100 times higher than previously available. The approach was successfully applied and tested in Cascais and Oeiras, two municipalities of the Lisbon Metropolitan Area, but since the approach is essentially data-based, accuracy of results is mostly

dependent on adequacy and quality of input data sets. However, based on these results it should be possible to implement this model for a number of additional municipalities in Portugal.

## 6. RECOMMENDATIONS FOR FURTHER STUDY

The following have been identified as potentially fruitful developments for the present work:

- Inclusion of some spatial indicator of density within dasymetric zones to improve upon an assumption of even distribution in these zones. One possibility would be to further experiment with information on land use and land cover from existing maps.
- Better modeling of people employed in distributed activities (e.g., security, cleaning, transportation, etc.) and modeling of people present in transportation networks, in hospitals and prisons, or in leisure and shopping activities.
- Better modeling may result in part from improving the LULC datasets by use of mixed classes or by using a main class and a secondary class to characterize mixed polygons.
- Better representation of population distribution during periods of commuting (8:00 to 10:00 AM and 5:00 to 8:00 PM).
- Using typical shopping profiles to allocate population to shopping centers during typical shopping hours.
- Using tourist census information to allocate tourists to hotels and other tourist attractions during typical times of sleeping and specific activities (beaches, museums, etc.).

- Additional research on the effect of spatial resolution (grid size) and gridding method on accuracy; experiment with gridding options to better account for right and left sides of streets.
- Further investigation on best methods to assess and portray quality measures for results.
- Explore statistical sources beyond census demographics to consider tourism influx in areas and periods where that activity is significant.
- Increased temporal segmentations of population distribution, so as to represent differences on a weekly basis (workdays vs. weekend) or on a seasonal basis (winter vs. summer).
- Use of model results to derive parameter weights for rapid estimation of daytime population distribution at regional scale with high spatial resolution.



## 7. BIBLIOGRAPHY

- Balk D. and G. Yetman, 2004. *The Global Distribution of Population: Evaluating the Gains in Resolution Refinement*. CIESIN, Columbia University, NY, USA.
- Balk, D. L., U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay and A. Nelson, 2006. In Hay, S.I., Graham, A.J. and Rogers, D.J. (eds), *Global mapping of infectious diseases: methods, examples and emerging applications. Advances in Parasitology*, volume 62. (London: Academic Press) pp. 119-156.
- Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Jerome Dobson, 2002. LandScan: Locating People is What Matters. *Geoinformatics* Vol. 5, No. 2, pp. 34-37.
- Bossard, M., Feranec, J., and Otahel, J., 2000. CORINE Land Cover Technical Guide – Addendum 2000. *Technical report No 40*. Copenhagen (EEA).
- Burt, J. E., and G. Barber, 1996. *Elementary Statistics for Geographers*. The Guilford Press, New York.
- Caetano, M., F. Mata, and S. Freire, 2006. Accuracy assessment of the Portuguese CORINE Land Cover map. In *Global Developments in Environmental Earth Observation from Space* (A. Marçal, Ed.), Millpress, Rotterdam, pp. 459-467.
- Chen, K, McAneney J., Blong R., Leigh R., Hunter L., and Magill C., 2004. Defining area at risk and its effect in catastrophe loss estimation: a dasymetric mapping approach. *Applied Geography*, 24:97-111.
- Chrisman, N., 2002. *Exploring Geographic Information Systems*, 2nd Ed., New York: John Wiley and Sons, 305 p.
- Clark, C., 1951. Urban Population Densities. *Journal of the Royal Statistical Society*, vol. 114 (Series A), no. 4, pp. 490-496.
- Deichmann, U., 1996. A review of spatial population database design and modeling. *Paper prepared for the UNEP/CGIAR Initiative on the Use of GIS in Agricultural Research*, National Center for Geographic Information and Analysis (NCGIA), University of California, Santa Barbara (UCSB), Santa Barbara, USA.

- Deichmann, Uwe, Deborah Balk and Gregory Yetman, Oct. 2001. Transforming Population Data for Interdisciplinary Usages: From Census to Grid. *NASA Socioeconomic Data and Applications Center (SEDAC)*, Columbia University, Palisades, NY, USA. Working Paper available on-line at: <http://sedac.ciesin.columbia.edu/plue/gpw/GPWdocumentation.pdf>.
- DeMers, M. N., 1997. *Fundamentals of Geographic Information Systems*. New York: John Wiley & Sons.
- DGEEP, 2001. *Workplaces and employment database*. Unpublished data, acquired on December 2006.
- Dobson, J. E., 2000. LandScan 1998 Provides Global Population Data at High Spatial Resolution. *GeoWorld*, Vol. 13, No.1, pp. 24-25.
- Dobson, J. E., 2002. War is God's Way of Teaching GIS. *Proceedings of the Merrill Conference on Science at a Time of National Emergency*. (<http://www.merrill.ku.edu/publications/2002whitepaper/dobson.html>).
- Dobson, J. E., 2003. Estimating Populations at Risk. Chapter 5.5 in *Geographical Dimensions of Terrorism* (Susan L. Cutter, Douglas B. Richardson, and Thomas J. Wilbanks, Ed.) Routledge: New York and London, pp. 161-167.
- Dobson, J. E., 2007. In Harm's Way: Estimating Populations at Risk. Technical paper in NRC [National Research Council]. Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises. Washington, D.C.: National Academy Press, p. 161-166.
- Dobson, J. E., E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley, 2000. A Global Population Database for Estimating Population at Risk. *Photogrammetric Engineering & Remote Sensing*, 66(7), pp. 849-857.
- Dobson, J. E., E. A. Bright, P. R. Coleman, and B. L. Bhaduri. 2003. LandScan2000: A New Global Population Geography, Chapter 15 in V. Mesev (ed.) *Remotely-Sensed Cities*, London: Taylor & Francis, pp. 267-279.
- Eicher, C. L., and Brewer, C. A., 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28, 125–138.

- FEMA, 2004. *Using HAZUS-MH for Risk Assessment*. Technical Manual, FEMA 433. Washington, DC: Federal Emergency Management Agency. Available at [www.fema.gov/HAZUS](http://www.fema.gov/HAZUS).
- Fisher, P., and M. Langford, 1995. Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27:211-24.
- Fisher, P. F., and Langford, M., 1996. Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. *The Professional Geographer*, 48: 299–309.
- Flowerdew, R., and M. Green, 1992. Developments in areal interpolation methods and GIS. *The Annals of Regional Science*, 26, 67–78.
- Freire, S., 2007a. Onde estão as pessoas quando não estão em casa? Modelação em SIG das distribuições diurnas e nocturnas da população de Cascais e Oeiras para avaliação de risco e apoio a emergências. *Actas da Conferência STIG – Saúde e Tecnologias de Informação Geográfica*, Lisboa, Portugal, 31 Maio-1 Junho, 14 pp.
- Freire, S., 2007b. O projecto DemoCarto: modelação em SIG da distribuição espacial e temporal da população de Cascais e Oeiras com alta resolução. *Actas do VI Congresso da Geografia Portuguesa*, Lisboa, Portugal, 17-20 Outubro, 23 pp.
- Freire, S. and M. Caetano, 2005. Assessment of Land Cover Change in Portugal from 1985 to 2000 Using Landscape Metrics and GIS. *Proceedings of GIS Planet 2005*, Estoril, May 30 – June 2, 20 pp.
- Gallego, J., and S. Peedell, 2001. Using CORINE Land Cover to map population density. *Towards Agri-environmental indicators, in Topic report 6/2001*, European Environment Agency, Copenhagen, pp. 94-105.
- Goldewijk, K. and J. J. Battjes, 1997. A hundred year (1890 - 1990) database for integrated environmental assessments (HYDE, version 1.1). *Report no. 422514002*, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands.
- Goodchild, M. F, and N. S. Lam, 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-processing*, 1: 297-312.

- Goodchild, M. F., L. Anselin, and U. Deichmann, 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25: 383-97.
- Haaland, C., and Heath, M., 1974. Mapping of Population Density. *Demography*, vol. 11, no. 2, pp. 321-336.
- Harvey, J. T., 2002a. Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing*, 23(10): 2071-2095.
- Harvey, J. T., 2002b. Population estimation models based on individual TM pixels. *Photogrammetric Engineering & Remote Sensing*, 68(11), pp. 1181-1192.
- INE (Instituto Nacional de Estatística), 2001. *Recenseamento Geral da População e da Habitação*. Lisboa.
- INE (Instituto Nacional de Estatística), 2003a. Movimentos Pendulares e Organização do Território Metropolitano: Área Metropolitana de Lisboa e Área Metropolitana do Porto 1991-2001. Lisboa.
- INE (Instituto Nacional de Estatística), 2003b. *Estatísticas Demográficas 2002*. Lisboa.
- Julião, R. P., 2003. Restructuring the Geographical Information Production and Dissemination at National Level – The Experience of Portugal, *Cambridge Conference Proceedings*, Ordnance Survey UK.
- (ORNL) Oak Ridge National Laboratory, 2007. *LandScan<sup>TM</sup> Global Population Database*. Oak Ridge, TN: Oak Ridge National Laboratory. Available at <http://www.ornl.gov/landscan/>
- Langford, M., 2006. Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, 30: 161–180.
- Langford, M., 2007. Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems*, 31: 19–32.
- Langford, M., Maguire, D. J., and D. J. Unwin, 1991. The Areal Interpolation Problem: Estimating Population Using Remote Sensing in a GIS Framework. In *Handling Geographical Information: Methodology and*

*Potential Applications*, Masser, I. and M. Blakemore (Eds.), New York, NY: Wiley, 55-77.

Leddy, R., 1994. Small area populations for the United States. *Presented at the Annual Meeting of the Association of American Geographers*, San Francisco, CA.

Lo, C.P. 2001. Modeling the population of china using DMSP operational linescan system nighttime data. *Photogrammetric Engineering & Remote Sensing*, 67(9): 1037-1047.

Lu, D., Weng, Q. and G. Li. 2006. Residential population estimation using remote sensing derived impervious surface. *International Journal of Remote Sensing*, 27(16): 3553-3570.

Maantay, J. A., Maroko, A., and Herrmann, C., 2007. Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS), *Cartography and Geographic Information Science*, special issue: Cartography 2007: Reflections, Status, and Prediction. Vol. 34, No. 2, pp. 77-102.

Martin, D., 1996. An Assessment of Surface and Zonal Models of Population. *International Journal of Geographical Information Systems*, 10(8):973-989.

Martin, D., and I. Bracken, 1991. Techniques for modelling population-related raster databases. *Environment and Planning A*, 23:1069-75.

McCleary, G. F., Jr. 1969. *The dasymetric method in the thematic cartography*. Unpublished Ph.D. dissertation, University of Wisconsin at Madison.

McCleary, G. F., Jr. 1984. Cartography, geography, and the dasymetric method. In: *Proceeding, 12<sup>th</sup> Conference of International Cartographic Association*, August 6-13, Perth, Australia. 1:599-610.

McPherson, T. N. and M. J. Brown, 2003. Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities. *Preprints: 84th AMS Annual Meeting*, AMS, Seattle, WA.

McPherson, T. N., A. Ivey and M. J. Brown, 2004. Determination of the spatial and temporal distribution of population for air toxics exposure assessments. *5th AMS Urban Env. Conf.*, Vancouver, B.C., 11 pp.

- McPherson, T. N., J. Rush, H. Khalsa, A. Ivey, and M. J. Brown, 2006. A Day-Night Population Exchange Model for Better Exposure and Consequence Management Assessments. *86th AMS Annual Meeting*, Atlanta, GA., 6 pp.
- Mennis, J., 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55:31-42.
- Mennis, J. and T. Hultgren, 2006a. Intelligent dasymetric mapping and its comparison to other areal interpolation techniques. *Proceedings of AutoCarto 2006*, June 26-28, Vancouver, WA.
- Mennis, J. and T. Hultgren, 2006b. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, vol. 33 (3):179-194.
- Néry, F., P. Monterroso, A. Santos, and J. Matos, 2007. Interpolação Zonal de Estatísticas Sócio-económicas. *Actas V Conferência Nacional de Cartografia e Geodesia*, Ed. Lidel: 89-99.
- NRC (National Research Council), 2007. *Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises*. Report by the National Academy of Sciences, Washington, D.C.: National Academy Press, 264 p.
- Oliveira, C. S., F. Mota de Sá, and M. A. Ferreira, 2005. Application of two different vulnerability Methodologies to Assess Seismic Scenarios in Lisbon. *Proceedings of 250th Anniversary of the 1755 Lisbon Earthquake*. November 1-4, Lisbon, Portugal.
- Openshaw, S., 1983. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, vol. 38. Norwich: Geobooks.
- Openshaw, S., 1984. Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16:17-31.
- Poulsen, E. and Kennedy, L., 2004. Using Dasymetric Mapping for Spatially Aggregated Crime Data. *Journal of Quantitative Criminology*, 20:243-262.
- Pozzi, F., Small, C. and Yetman, G., 2003. Modeling the distribution of human population with night-time satellite imagery and gridded population of the world. *Earth Observation Magazine* 12. Available at: [http://www.eomonline.com/Common/Archives/2003jun/03jun\\_humanpop.html](http://www.eomonline.com/Common/Archives/2003jun/03jun_humanpop.html)

- Rabbani, S. K., 2007. *Counting populations-at-risk: co-verification of LandScan database and building occupancy coefficients*. Unpublished M. A. thesis, University of Kansas.
- Reibel, M., and Bufalino, M., 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37, 127–139.
- Robinson, A., J. Morrison, P. Muehrcke, A. Kimerling, and S. Guptill, 1995. *Elements of Cartography*, 5<sup>th</sup> ed., New York: John Wiley and Sons.
- Silverman, B. W., 1986. *Density estimation for statistics and data analysis*. New York: Chapman and Hall.
- Sleeter, R., 2004. Dasymetric mapping techniques for the San Francisco Bay region, California: *Urban and Regional Information Systems Association*, Annual Conference, Proceedings, Reno, Nev., November 7–10, 2004.
- Sleeter, R., and N. Wood, 2006. Estimating daytime and nighttime population density for coastal communities in Oregon: *Urban and Regional Information Systems Association*, Annual Conference, Proceedings, Vancouver, BC, September 26-29, 2006.
- Sutton, P., 1997. Modeling Population Density with Nighttime Satellite Imagery and GIS. *Computers, Environment, and Urban Systems*, 21(3/4): 227-244.
- Sutton, P., C. Elvidge, and T. Obremski, 2003. Building and evaluating models to estimate ambient population density. *Photogrammetric Engineering & Remote Sensing*, 69(5): 545-553.
- Sutton, P., D. Roberts, C. D. Elvidge, and H. Meij, 1997. A comparison of nighttime satellite imagery and population density for the continental United States. *Photogrammetric Engineering and Remote Sensing*, 63, 1303–1313.
- Sutton, P., D. Roberts, C. D. Elvidge, and K. Baugh, 2001. Census from Heaven: an estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing*, 22(16): 3061-3076.
- Sweitzer, J. and Langaas, S., 1995. Modelling population density in the Baltic Sea States using the Digital Chart of the World and other small scale data sets. In Gudelis, V. Povilanskas, R. and Roepstorff, A. (eds.). *Coastal*

*Conservation and Management in the Baltic Region*. Proceedings of the EUCC -WWF Conference, 2-8 May 1994, Riga - Klaipeda - Kaliningrad, pages 257-267.

Tobler, W. R., 1979. Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*, Vol. 74, No. 367 (Sep., 1979), pp. 530-535.

Tobler, W. R., U. Deichmann, J. Gottsegen, and K. Maloy, 1995. The Global Demography Project. *Technical Report No. 95-6*. National Center for Geographic Information and Analysis. UCSB. Santa Barbara, CA, 75pp.

Vidal, C., J. Gallego, and M. Kayadjanian, 2001. Geographical use of statistical data. *Topic report 6/2001*, European Environment Agency, Copenhagen, pp. 11-24.

Wright, J. K., 1936. A Method of Mapping Densities of Population, *The Geographical Review*, 26(1):103-110.

Wu, C., and Murray, A., 2007. Population Estimation Using Landsat Enhanced Thematic Mapper Imagery. *Geographical Analysis*, 39: 26-43.

Wu, S-S., Qiu, X., and Wang, L., 2005. Population estimation methods in GIS and remote sensing: a review. *GIScience and Remote Sensing*, 42: 80-96.

Zandvliet, R., and M. Dijst, 2004. Short-term Shifts in Population Distribution - Unraveling the Diurnal Mobility of Daytime Population in the Netherlands. *Paper presented at the 26th IATUR Conference, 27-29 October, Rome, Italy*.